CSE-519: Data Science Fundamentals

PROJECT REPORT

Trade Values Analysis in Baseball and Other Sports

ANKIT AGRAHARI (110946823)
ARPITH VARGHESE (110872883)
PIYUSH SHYAM BANGINWAR(110944214)

Problem Statement:

Players are routinely traded between teams, for other players of presumably equal value. Develop a system to estimate the future value of two sides in a trade.

Objectives:

Our analysis and prediction methods will primarily be focussed on Baseball trading which can later be extended to analyse other sports. We intent to evaluate the following:

1) Perform trade value analysis by using player performance metrics after the trade to be the face value of a side to decide which side got a better deal.
2) Perform trade value analysis by assuming that trades are of equal value at the time of trade and finding parameters that make them of equal value. These parameters are compared with the actual values to measure if the trade really benefitted both sides equally.
3) Use the score from trade value analysis to come up with a prediction technique to estimate the lopsidedness of a trade.

Literature Review:

Baseball : Baseball is a game played between two teams, each composed of nine players. The game consist of nine innings, and the team with the greater number of runs at the end of the game wins. The goal of the game is to score more points (runs) than the other team.

Major Teams and Leagues : Major League Baseball (MLB) is a professional baseball organization consist of two component leagues National League and American League. A total of 30 team (29 America and 1 Canada) play in National League and American League with 15 in each League.

Trading in Baseball : Baseball transactions are changes made to the roster of a team during or after the season. They may include waiving, releasing, and trading players, as well as assigning players to a lower league teams. There are certain deadlines for the type of trade being carried out during a season. For instance, the July 31 non-waiver trade deadline and Aug 31 the waiver deadline.

Wins Above Replacement: or Wins Above Replacement Player, commonly abbreviated to WAR, is a non-standardized sabermetric baseball statistic developed to sum up "a player's total

contributions to his team". A player's WAR value is claimed to be the number of additional wins his team has achieved above the number of expected team wins if that player were substituted by a replacement-level player.
WAR is calculated separately for pitchers and for positional players

1) WAR for positional players: WAR for positional players have six components: Batting runs, Base running runs,Fielding runs, Runs added or lost due to grounding into double plays. Positional adjustment runs, Replacement level runs
2) WAR for pitchers: Pitching WAR calculation requires only overall Runs allowed (both earned and unearned) and Innings Pitched.

Dataset:

We have the following datasets available with us:

Dataset	**Timeline**	**No. of Entries**	**Major Columns/Features**	**Data Source**
Transactions	1873-2015	85144	Date From-Team To-Team Players Involved	http://retrosheet.org/
Player Statistics - Batting	1871-2014	101332	games, runs, hits, doubles, triples, home runs, strikeouts, runs batted in etc	http://www.seanlahman.com/baseball-archive/statistics/
Player Statistics - Pitching	1871-2014	44139	Games, Wins, Loss, Grand Slam,ERA, WHIP Strikeouts,Saves	http://www.seanlahman.com/baseball-archive/statistics/
Player Statistics - Fielding	1871-2014	170526	Fielding Position,Games Started,	http://www.seanlahman.com/baseball-archive/statistics/
Teams Data	1871-2015	2805	Team Name,Wins, Loss, Rank,	http://www.seanlahman.com/baseball-archive/statistics/

Player Salaries	1985-2015	25575	PlayerId, Salary	http://www.seanlahman.com/baseball-archive/statistics/
Pitchstats	1871-2016	102138	Team, Year, RunsBat, RunsInfield, WAA, WAR, isPitcher	http://www.baseball-reference.com/about/
Batstats	1871-2016	102138	Team, Year, RunsBat, RunsInfield, WAA, WAR, isPitcher	http://www.baseball-reference.com/about/

Trade Value Analysis:

Analysing the trades for their fairness will be an important aspect of the project as our data is not labelled. The analysis we come up with will be used to label/score the data for their fairness and any prediction model being developed will take this as the benchmark to train itself.

We perform the analysis in two major ways:

- Cumulative WAR Value addition on both sides over N years after trade.
- Assume trades to be equal value at the time of trade to get decay ratios and use them to analyse with actual values.

Both are explained in detail below:

a) Using Cumulative WAR values

This method will try to evaluate a trade for its fairness by calculating its lopsided nature (if present) after a certain amount of time from the trade date. The calculation makes use of the Wins Above Replacement (WAR) metric mentioned earlier, which gives us the performance measurement of a player(at age X) for a year. And we evaluate how good a player is by comparing his performance with the average performance of any player at that age.
The exact method works the following way:

1. Calculate total value of the trade for each side over a period of N years and return the absolute difference between these two values as the trade value difference.
2. Values near zero correspond to fair trades whereas those away from zero denote the lopsided trades.
3. The value of each side is calculated by summing the Value of each player involved on their side of the deal. Here, the player value is calculated by finding their cumulative value over N years after the trade. The value for a particular year is the difference between his actual and expected performance for that year.

4. Actual Performance of a player (age X) in a year = WAR of the player for that year.
5. Expected Performance of a player of age X = Average WAR at age X.

Assumptions Taken:

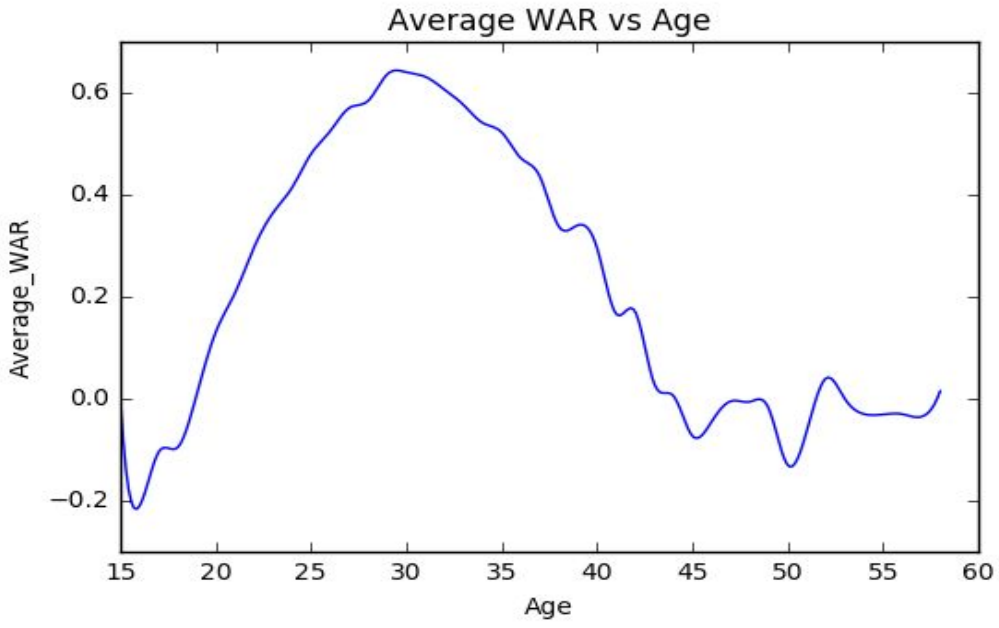
1. For simplicity, we take N = 15 years.
2. For any year when a player is not playing (maybe got retired or injured, or the data not available for that year) we assume his contribution/value to be zero.
3. Over the course of these 15 years, it is possible that the player got traded to another team, in this case, we assume that the trade must have been a equal value trade and the team got another player of almost equal calibre in exchange and hence ignoring the trade in calculation.
4. We take the average performance of a player at age X as the average of WAR values of all the players at age X in our database. (Results shown below)
5. We ignore cash transactions involved as part of a trade.

$$V_{0,1} = \sum_{i=0}^{n}\left(\sum_{j=0}^{15}(WAR_{i,j} - Avg\ WAR_j)\right)$$

Here, $n = No.\ of\ players\ involved\ in\ team\ 0/1's\ side$

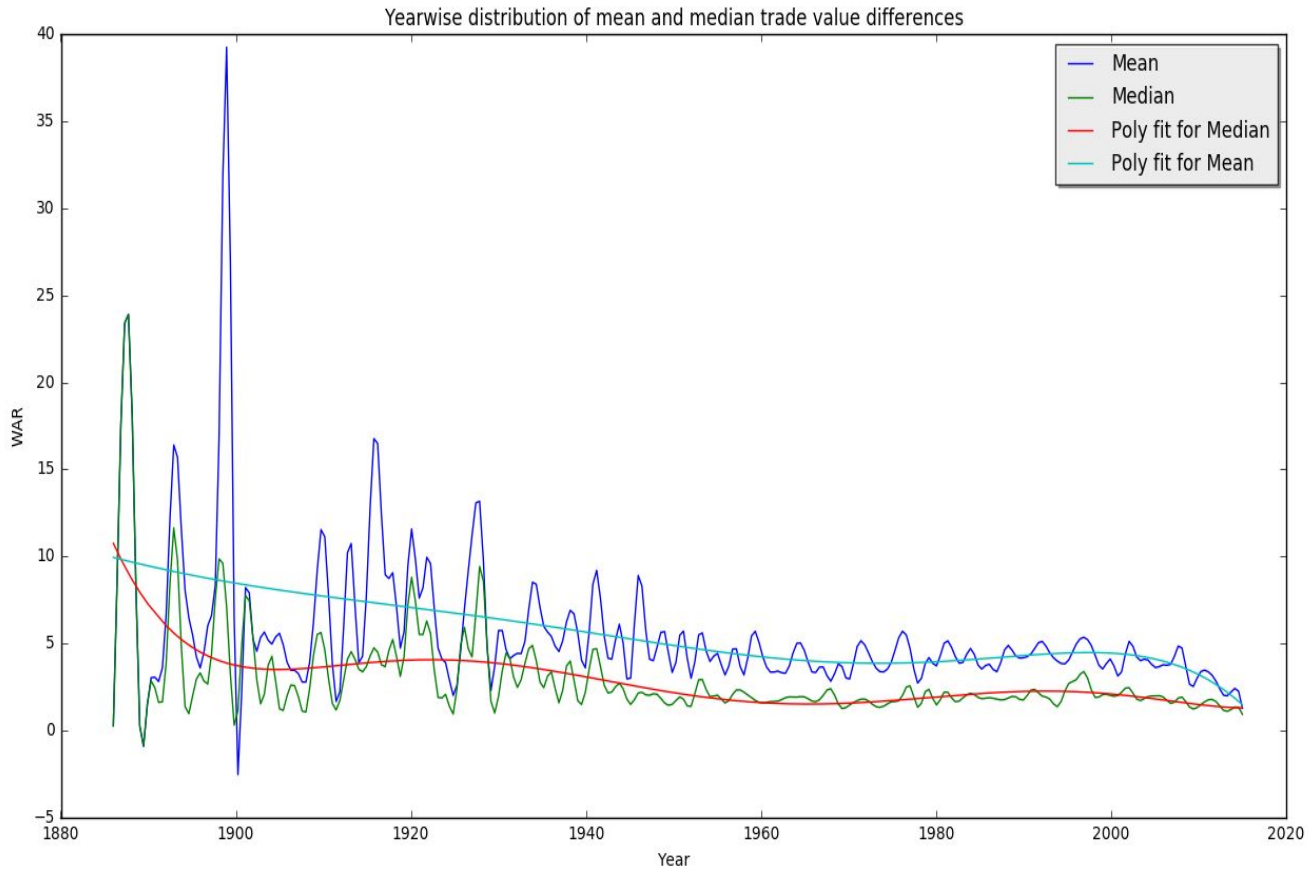
$$WAR_{i,j} = WAR\ of\ Player\ i\ during\ (trade + j)^{th}\ year$$

$$Avg.\ WAR_j = Average\ WAR\ for\ age\ x,\ x = age\ of\ player\ i\ during\ (trade\ date + j)$$



Results from Analysis:

Year wise distribution of mean and median trade value differences:



The above graph shows that the number of lopsided trades is decreasing year on year probably because teams have become more cautious and they perform advanced player performance evaluation before finalizing a trade.

Top 10 results as per this analysis:

Rank	Date	Team1	Team2	Player/s from Team 1	Player/s from Team 2	Who got the better deal
1	December 8, 1899	Louisville Colonels	Pittsburgh Pirates	Honus Wagner, Fred Clarke, Bert Cunningham, Mike Kelley, Tacks Latimer, Tommy Leach, Tom Messitt, Deacon Phillippe, Claude Ritchey,Rube Waddell	Jack Chesbro, George Fox, Art Madison, John O'Brien	Pittsburgh Pirates
2	April 9, 1916	Boston Red Sox	Cleveland Indians	Tris Speaker	Sad Sam Jones, Fred Thomas	Cleveland Indians
3	August 30,1990	Houston Astros	Boston Red Sox	Larry Andersen	Jeff Bagwell	Boston Red Sox
4	December 5, 1984	Oakland Athletics	New York Yankees	Rickey Henderson,Bert Bradley	Stan Javier,Jay Howell,Jose Rijo,Eric Plunk,Tim Birtsas	New York Yankees
5	February 2, 1893	Cleveland Spiders	New York Giants	George Davis	Buck Ewing	New York Giants
6	December 5, 1998	Texas Rangers	Chicago Cubs	Mitch Williams,Paul Kilgus,Steve Wilson,Curt Wilkerson	Rafael Palmeiro,Jamie Moyer,Drew Hall	Texas Rangers
7	January 27, 1982	Chicago Cubs	Philadelphia Phillies	Ivan de Jesus	Larry Bowa,Ryne Sandberg	Chicago Cubs
8	July 23, 1910	Philadelphia Athletics	Cleveland Naps	Morrie Rath,Shoeless Joe Jackson	Bris Lord	Cleveland Naps
9	December 10, 1991	Houston Astros	Cleveland Indians	Kenny Lofton,Dave Rohde	Willie Blair,Ed Taubensee	Cleveland Indians
10	November 29, 1971	Cincinnati Reds	Houston Astros	Lee May,Tommy Helms,Jimmy Stewart	Joe Morgan,Denis Menke,Jack Billingham,Cesar Geronimo,Ed Armbrister	Cincinnati Reds

b) Assuming Trades are of Equal Value:

It is a rule of business that a trade can happen between two parties only if both assume that they are getting a fair deal. This is applicable in our scenario too, where baseball teams while trying to make a trade will try to maximize the value they get out of the deal and hence the trades will ideally be of equal value. This is an important assumption we are making to further base our analysis and prediction on. It will work the following way:

- In the previous analysis we had assumed that the projected performance of a player will be equal to the average player performance curve.
- In order to get a more accurate estimate of player projection , we project the performance of a player over a span of 15 years based on his past performance and age.
- We then add up the values for each player from a team.

$$x = Mean(Actual\ WAR_i - Average\ WAR_i),\ over\ the\ player's\ career\ before\ the\ trade$$

$$where,\ i = age\ of\ player$$

$$Expected\ WAR_i = Average\ WAR_i + x,\ where\ i = Age\ of\ Player$$

- We then combine these values with a decay rate similar to the amount in Compound Interest and try to find rates for both teams such that the trade values at the time of trade are as close as possible.

$$At\ time\ of\ trade,\ trade\ values(F1, F2)\ are\ assumed\ equal\ over\ the\ course\ of\ 15\ years.$$

$$F_1 = F_2$$

Decay Rate:

$$Trade\ Value\ Diff = \sum(Proj\ WAR_{1j} * (1 - r1/100)^j) - \sum(Proj\ WAR_{2j} * (1 - r2/100)^j)$$

$$Here,\ 1, 2 = Teams;\ r1, r2 = decay\ rates\ for\ each\ team;\ j = no.\ of\ years\ after\ trade$$

- Our objective will be to make the Trade Value Diff to be as close to zero as possible, and we know the values of all parameters other than r1, r2.
- To find these rates of decay, we use an optimization technique, **Truncated Newton Method** which are a family of optimization algorithms designed for optimizing non-linear functions with large numbers of independent variables.
- A Truncated Newton method consists of repeated application of an iterative optimization algorithm to approximately solve Newton's equations, to determine an update to the function's parameters.
    - We used to *scipy.optimize* library of Python to specify this method (TNC) for our optimization.
- Once we have these decay rates calculated, we compute the trade value for 15 years from the trade date with actual values of WAR using the same decay rates that was computed earlier for each team, and use these computed values to judge which team got the better side of the deal.

$$Trade\ Value\ Diff\ =\ \sum(Act\ WAR_{1j}\ *\ (1-r1/100)^{j})\ -\ \sum(Act\ WAR_{2j}\ *\ (1-r2/100)^{j})$$

The Trade Value Diff here denotes the lopsidedness of the trade.

Supporting Logic: The logic behind this procedure is that: while performing the trade, both sides assumed the deal to be fair based on some factors associated with their own team and by projecting the value gain for going ahead with the trade. These are nothing but assumptions that the teams are making. Now, these assumptions might not turn out to be true in reality and the lopsidedness of the trade is nothing but the deviation from these assumptions in favor of one side or the other.

This method tries to take these things into consideration by calculating decay ratios at trade time and applying them to actual values, to calculate the deviation from the supposed zero/equal value.

- Also, unlike the earlier method, we also account for cash transactions this time. To do this, we found the equivalent of cash in terms of WAR using the salaries tables. All cash values were inflation adjusted before finding their equivalent WAR value. The procedure followed was:
    - Adjust all Salaries to 2016 CPI and take their average.
    - Take the average WAR for the same set.

- Equate the two to find the relation

  From our analysis we found that:

  1 WAR = $ 3,928,257

These values were substituted wherever applicable.

Results from Analysis:

Top 10 results as per this analysis:

Rank	Date	Team1	Team2	Players Team1	Player Team2	Trade Value Diff
1	December 14, 1949	New York Giants	Boston Braves	Sid Gordon,Buddy Kerr,Willard Marshall,Red Webb	Eddie Stanky,Al Dark	8.103336
2	December 6, 1959	Cleveland Indians	Chicago White Sox	Minnie Minoso,Dick Brown,Don Ferrarese,Jake Striker	John Romano,Bubba Phillips,Norm Cash	7.831614
3	January 14, 1963	Baltimore Orioles	Chicago White Sox	Hoyt Wilhelm,Dave Nicholson,Pete Ward,Ron Hansen	Luis Aparicio,Al Smith	7.402284
4	11 December, 1959	New York Yankees	Kansas City Athletics	Don Larsen,Hank Bauer,Norm Siebern,Marv Throneberry	Roger Maris,Joe DeMaestri,Kent Hadley	7.368413
5	September 30,1946	Pittsburgh Pirates	Boston Braves	Bob Elliott,Hank Camelli	Billy Herman,Elmer Singleton,Stan	7.213203

					Wentzel,Whitey Wietelmann	
6	December 12,1999	San Diego Padres	Atlanta Braves	Wally Joyner,Reggie Sanders,Quilvio Veras	Bret Boone,Ryan Klesko,Jason Shiell	7.159809
7	December 26,1904	Boston Americans	St. Louis Browns	George Stone	Jesse Burkett	7.08906
8	October 29,1928	New York Giants	Philadelphia Phillies	Lefty O'Doul	Freddy Leach	6.958705
9	November 18, 1997	Thornton Baseball Association	Philadelphia Phillies	Bobby Abreu	Kevin Stocker	6.815524
10	November 7, 1928	Chicago Cubs	Boston Braves	Socks Seibold,Percy Jones,Lou Legett,Freddie Maguire,Bruce Cunningham	Rogers Hornsby	6.784525

Which one is better?

The first method is based on a naive approach of analysis where the trade value difference is calculated simply as the difference between the player performances after the trade. But it does not take in-to account the intent of the team or manager that goes in-to making the trade at first place.

A trade will be truly lopsided if the intent behind making the trade did not serve its purpose. The second method is trying to do that, where it assumes that the trades will in-fact become equal value over a course of time and measures the lopsidedness by calculating the deviation from this assumption.

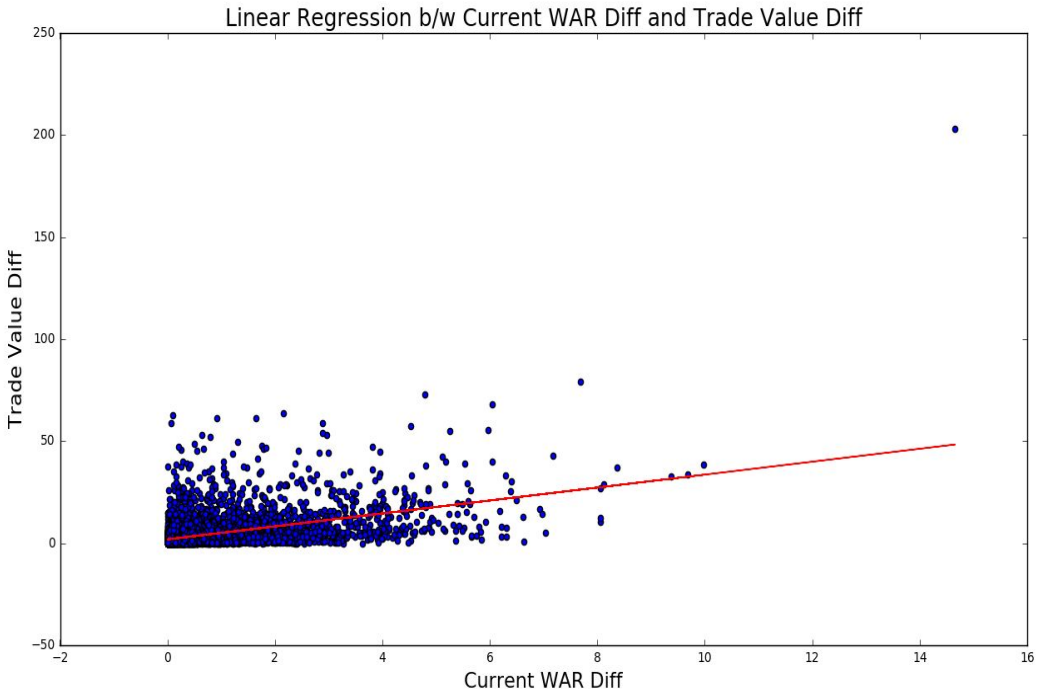
If seen from the perspective of a casual fan follower, Method 1 makes more sense, but in the purview of a team management, who are making the trade happen, Method 2 is more appropriate. As our basic assumption is that the trades are in-fact of equal value at time of trade, we will consider Method 2 to be the benchmark for our prediction models.

Prediction:

Baseline Model:

Our baseline model for prediction is based on the Cumulative WAR Technique/Method-1, we used to analyse the trades.

We calculate the value of a trade by measuring the total value on each side for the trade year, which is the sum of the values of the players involved on each side. The value of the player is taken same as in analysis, i.e; the Wins Above Replacement (WAR). In this case, we only consider the players WAR for the current year to simplify the model. We then evaluate this model with the values we got from the trade value analysis to measure its accuracy/correlation.



Linear Regression Mean Square Error: 6.52680325465

Correlation Coefficient: 0.469736332344
Correlation p-Value: 1.3179029177e-306

Why Baseline Model is Not That Great:

1. The model does not assume that the trades are of equal value during trade.
2. We assume that the WAR difference is the sole criteria of judging a trade and other factors are ignored.
3. It assumes the importance of a player remains consistent though his career (i.e; no decay rate).
4. We ignore cash transactions.

Advanced Model:

Our advanced model is based on the Equal Value Trade Analysis/Method-2 we performed, and will be taking those values to score our data and fit various regression models and further analyse their accuracy and error rates to choose the best one.

Using the available dataset, we tried to extract all features that could in any way affect the trade between two teams at the time of trade and find correlations between these features with the Trade Value Difference we calculated in our advanced/second analysis.

Feature Extraction:

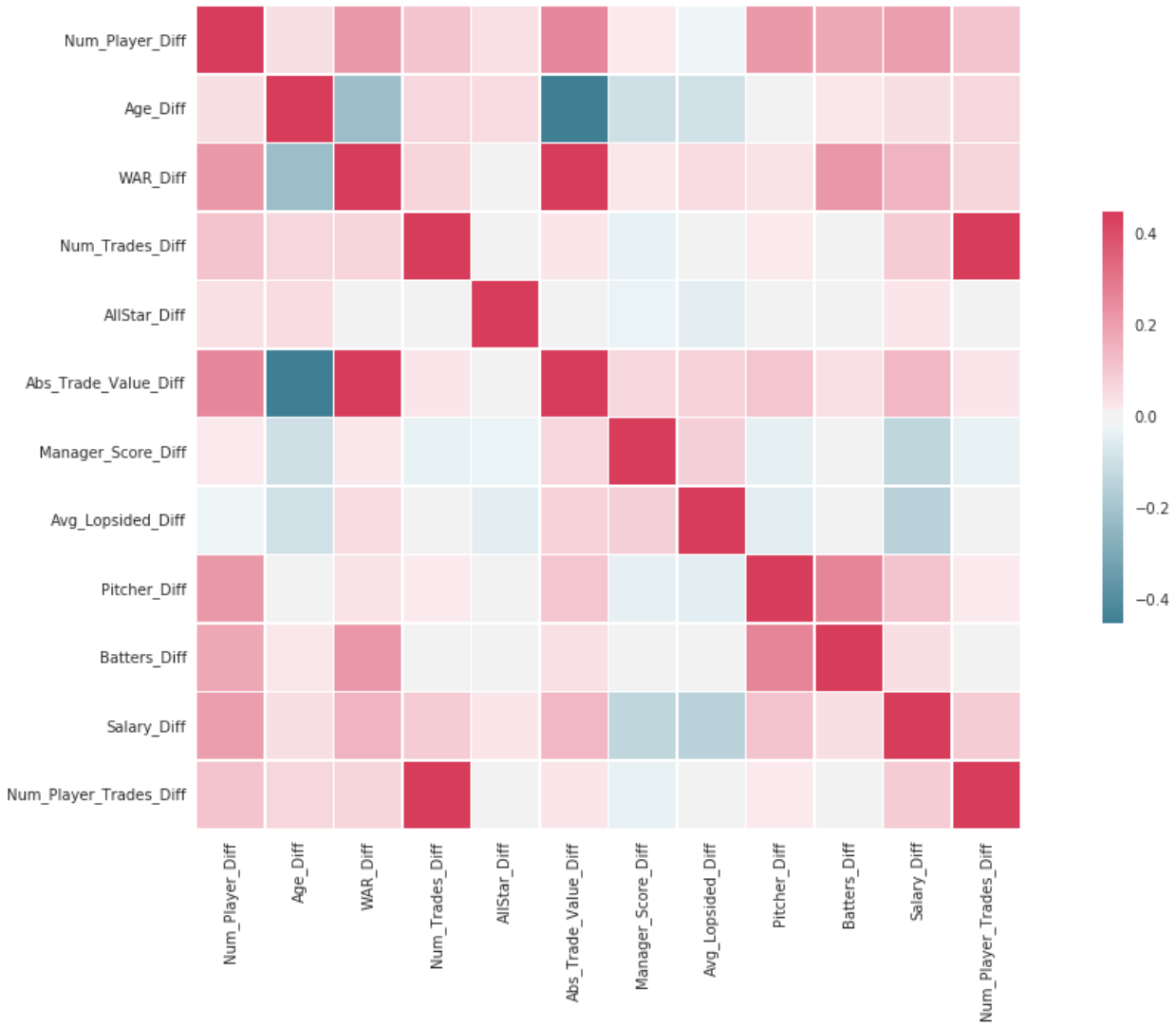
1) WAR Difference:
   Absolute difference of the WAR score at the time of trade for all players that are involved in a trade.
2) Salary Difference:
   Difference of salaries at the time of trade of all the players involved
3) Manager Performance:
   We calculate the manager score of both team involved in the trade.
   Manager score will be the manager's performance till the trading year.

   Manager Score = Sum of wins of all teams under the manager/ total game played by all team under the manager.

4) Age Difference:
   Cumulative age difference of players involved on each side of the trade.
5) Number of players: Number of players involved in a trade on each side.
6) All stars: Number of All-Star Players involved in the trade on each side.
7) Average past trade value: Average value of the trades made by both teams in the past.
8) No. Past trades of a player: Number of trades each player was involved in the past.
9) No. of Past trades of a team in a year : Number of past trades made by the team in the same year of the trade.
10) Batters Diff - Number of batters exchanged in the trade.
11) Pitchers Diff - Number of pitchers exchanged in the trade.

Feature Selection:

We selected the features from above extracted features that correlated the most with our analysed Trade Value Difference. The correlation matrix for which is:



Based on the correlation, the features we selected for our models are:

- WAR
- Number of players
- Age
- Rank
- Number of past trades for the team
- Salary
- Number of all star players involved
- Number of Pitchers
- Number of Batters
- Manager Score

Training & Testing Data Sets:

We split our data set in-to a 7: 3 ratio for providing training and testing samples. We also removed all trades for 2015 to run our prediction algorithm for that year.

Regression Models:

We used a number of regression models on our data to predict the trade value difference for a trade, indicating its lopsided nature. Following models were used:

1) Linear Regression
2) Ridge Regression
3) Lasso Regression
4) Decision Tree Regression
5) Random Forest Regression
6) Nearest Neighbour Regression
7) Polynomial Regression
8) Gradient Boosting Regression

Model Evaluation:

Regression metrics:

(Reference:http://scikit-learn.org/stable/modules/model_evaluation.html)

1) **Explained variance score :** Explained variation measures the proportion to which a mathematical model accounts for the variation (dispersion) of a given data set If $\hat{y}$ is the estimated target output, y the corresponding (correct) target output, and Var is Variance, the square of the standard deviation, then the explained variance is estimated as follows:

$$\texttt{explained\_variance}(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

The best possible score is 1.0, lower values are worse.

2) **Mean absolute error:** The mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. If $\hat{y}_i$ is the predicted value of

the i-th sample, and y_i is the corresponding true value, then the mean absolute error (MAE) estimated over n_{samples} is defined as:

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| .$$

3) **Mean squared error:** It is a risk metric corresponding to the expected value of the squared (quadratic) error loss or loss. If $\hat{y}_i$ is the predicted value of the i-th sample, and y_i is the corresponding true value, then the mean squared error (MSE) estimated over n_{samples} is defined as

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

4) **Median absolute error:** It is calculated by taking the median of all absolute differences between the target and the prediction. This value is robust to outliers

5) **R² score, the coefficient of determination:** It provides a measure of how well future samples are likely to be predicted by the model. Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0. If $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the corresponding true value, then the score R² estimated over n_{samples} is defined as

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1} (y_i - \bar{y})^2}$$

where $\bar{y} = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} y_i$

Results:

Model	Mean abs error	Median abs error	Root Mean Squared Error	Explained variance score	R2 score
Linear	0.41748	0.24071	0.74501	0.31453	0.31449
Ridge	0.41744	0.24080	0.74499	0.31457	0.31453
Lasso	0.55589	0.41423	0.88989	0.02245	0.02195
Decision Tree	0.42358	0.20106	0.84451	0.11916	0.11916
Random Forest	0.40258	0.20472	0.76223	0.28244	0.28243
K-Nearest Neighbour	0.50162	0.35991	0.87073	0.06425	0.06361
Polynomial	0.59973	0.31341	1.02488	-0.01695	-0.0186
Gradient Boosting	0.47523	0.1451	0.96979	0.08873	0.08787

We chose Random Forest Regressor as the best model as it performs well on all the error metrics.

The distribution of error for Random Forest Regressor iis shown in the below graph:

Probability of Winning/Losing:

The regression model gives us a predicted trade value given some input feature but we would also like to know the chances of a team winning or losing the trade, i.e. a probability value. We used Logistic Regression to estimate this probability as follows:
Based on the analysis of the lopsided trades, we know if team 1 got the better deal than team2. We assign labels such that whenever team 1 won, it is assigned a label 0 otherwise a label 1. This is used to train a logistic regression model on a subset of the input features used for regression, and predict the probability of a team winning the trade. The model produced results with an accuracy of 0.65 which is slightly better than random chance. Even though the model has a high error rate, we have still included the results for the sake of completeness.

Our prediction for the trades of 2015:

See Below:

Rank	Date	Team1	Team2	Players Team1	Player Team2	Trade Value Diff	Team 1 Win Percentage	Team 2 Win Percentage
1	2015-11-12	Atlanta Braves	Los Angeles Angels of Anaheim	Andrelton Simmons,Jose Briceno	Erick Aybar, Chris Ellis, Sean Newcomb	3.243647	71%	29%
2	2015-04-05	Atlanta Braves	San Diego Padres	Craig Kimbrel,Melvin Upton	Cameron Maybin,Carlos Quentin,Matt Wisler, Jordan Paroubeck	2.239327	50%	50%
3	2015-05-27	Los Angeles Dodgers	Atlanta Braves	Juan Uribe,Chris Withrow	Alberto Callaspo,Eric Stults,Ian Thomas,Juan Jaime	1.387574	22%	78%
4	2015-11-11	Minnesota Twins	New York Yankees	Aaron Hicks	J. R. Murphy	1.372885	60%	40%
5	2015-01-10	Thornton Baseball	Oakland Athletics	Ben Zobrist,Yunel Escobar	John Jaso, Boog Powell, Daniel Robertson	1.333948	51%	49%
6	2015-07-30	Miami Marlins	Atlanta Braves	Mat Latos,Mike Morse	Jose Peraza,Alex Wood,Bronson Arroyo,Jim Johns,Luis Avilan	1.256636	35%	65%
7	2015-01-19	Houston Astros	Chicago Cubs	Dexter Fowler	Luis Valbuena,Dan Straily	1.134766	56%	44%
8	2015-06-03	Arizona DiamondBacks	Seattle Mariners	Mark Trumbo,Vidal Nuno	Welington Castillo,Dominic Leone, Gabriel Guerrero, Jack Reinheimer	1.028551	65%	35%
9	2015-07-31	Philadelphia Phillies	Toronto Blues	Ben Revere	Jimmy Cordero and Alberto Tirado	1.015970	75%	25%
10	2015-08-07	Cleveland Indians	Atlanta Braves	Nick Swisher,Michael Bourn,	Chris Johnson	1.004355	52%	48%

References:

Dataset References:

- http://www.baseball-reference.com/about/
- http://www.seanlahman.com/baseball-archive/statistics/
- http://retrosheet.org/

Background updates:

- https://en.wikipedia.org/wiki/Wins_Above_Replacement

Advanced Model:

- https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/
- http://scikit-learn.org/stable/modules/model_evaluation.html