CMPE 333
Final Report

Prediction Methods to Determine if a Baseball Player is Worth Their Salary

Peter Lyons

11-17-2015

Contents

Background

Baseball is an incredible sport with one of the most advanced statistical collections and evaluation metric databases available. Most large sports networks stick to simple statistics: batting average (AVG), runs batted in (RBI), home runs (HR), and on-base percentage (OBP). However, these statistics barely scratch the surface of determining whether a baseball player is considered an effective player or not.

In baseball, players develop at very different rates than in hockey or basketball. There are two different routes to the draft: the college route or straight out of high school. By having two traditional routes to the draft, there is a larger span of ages drafted by MLB teams each year.

Players spend on average 4 years in the minors before making a major league roster, after which their club retains control for 6 years [1]. Having control does not mean the player is under contract, so salaries may rise while under club control. This process is known as salary arbitration, where a judge will award a salary based upon two offers submitted by the club and the player agent. The impact of having a wide range of ages to break into the league and clubs maintaining control for many years, means it is more difficult to determine if a player is worth their salary overall or for their age bracket.

In 2013 the MLB the median salary was $1.1 million whereas the mean salary was $3.4 million [2]. The reason for this discrepancy is that if a star player reaches free agency, they often sign contracts worth upwards of $100 million over a period of time. This situation results an inflation of the mean salary within the league.

Clustering Methods

From the first attempted plot between RBIs and player salary it can be seen there is no evident trend as the amount of RBIs remain fairly consistent from X salary to Y in Figure 1.

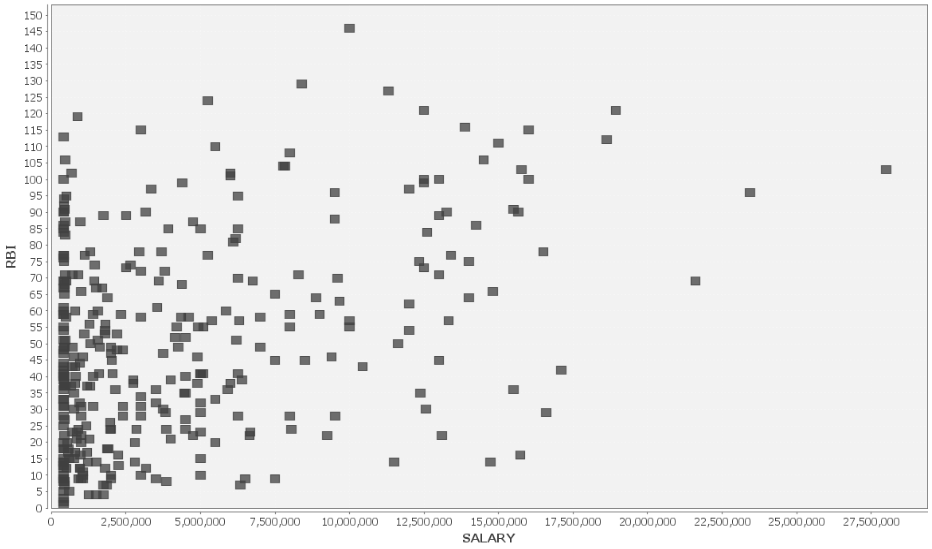


Figure 1: A plot depicting Runs Batted In vs. a player's annual salary

With the next simple visualization between OBP and Salary (Figure 2), results were again all over the place. But it can be seen that at higher salary ranges, the lowest OBP is much better than players with lower salaries.

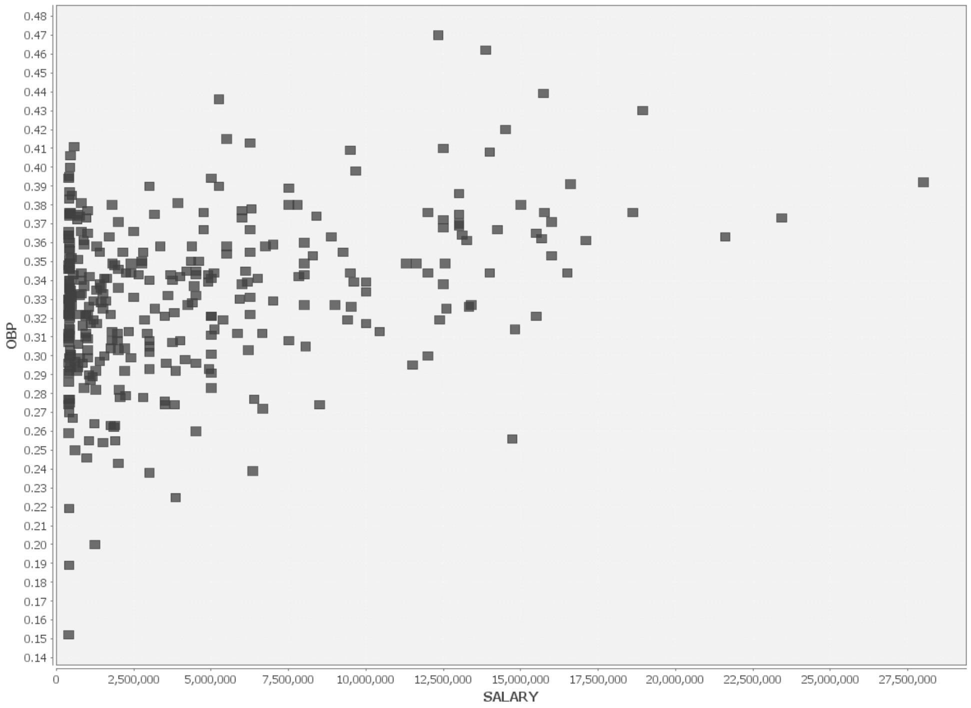


Figure 2: On Base Percentage compared to a player's annual salary

A histogram (not shown) was used on the sum of RBI and OBP normalized between 0.0 and 1.0. The reasoning behind this measurement was that power hitters often collect most of the RBIs on a team, yet there are many great players who play at the top of the order and get on base frequently. Therefore, the collective sum should provide a decent representation of what a great hitter can provide.

Unfortunately, no correlation was found between the normalized sum and the player salary. Reasons for this can be attributed to the lack of age data and defensive metrics given in the baseball set. Players must be broken down into age brackets, or tiers because of the way MLB structures contracts. Furthermore, half the game of baseball is defense. Without proper defensive metrics included in the dataset, the analysis of a player’s worth becomes solely offensively focused, and therefore is not a true representation of the player.

To further break down the data, a minimum of 100 games played was chosen to reflect whether a player was worth their salary for the season. Even though a player may be injured and on the disabled list, the team still needs to pay their salary.

To account for any missed games, a new column was created with RBI per game average. Again, the new RBI/Game and OPS (on-base plus slugging) were normalized between 0.0 and 1.0. The results were plotted and there was no clear trend showing that higher salary players were better.

Since, OPS does not take into account the quality of the on-base production, a new stat was used. Based on available data in the dataset, wOBA (weighted on base average) was calculated.

$$wOBA = \frac{0.72 * UBB + 0.75 * HBP + 0.9 * 1B \ + 0.92 * RBOE + 1.24 * 2B + 1.56 * 3B + \ 1.95 * HR}{PA}$$

wOBA is one of the most common advanced metrics used to determine the offensive contribution of a player while factoring in quality of the production [3]. The wOBA plot in Figure 3 shows a great linear bottom line, where a line fit may also be able to determine if a player is worth their salary.

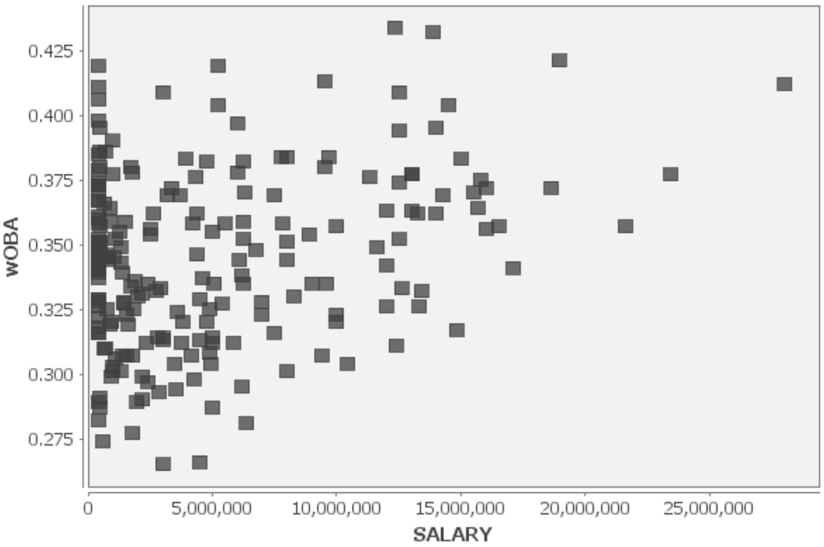


Figure 3: Players wOBA vs. salary

Additionally, value over replacement player (VORP) was also considered to be an important statistic when determining whether a player is considered above average at their position. Average VORP for the 2013 MLB season was around 20, so theoretically if a player had a VORP over 20, they would be above-average and therefore worth their salary.

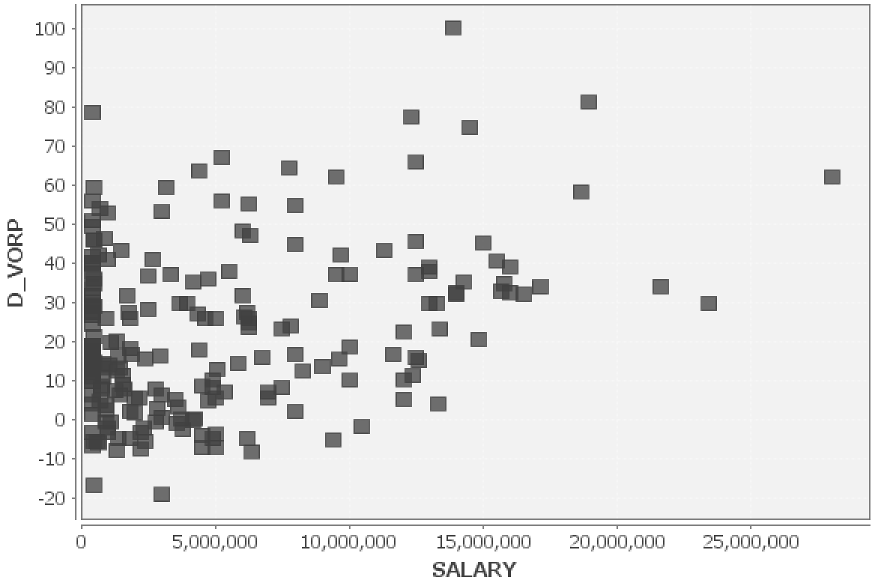


Figure 4: D_VORP vs. Salary

It can be seen in Figure 4 that there is a clear divide of players above the VORP threshold of 20, and a better visual of player value is starting to emerge.

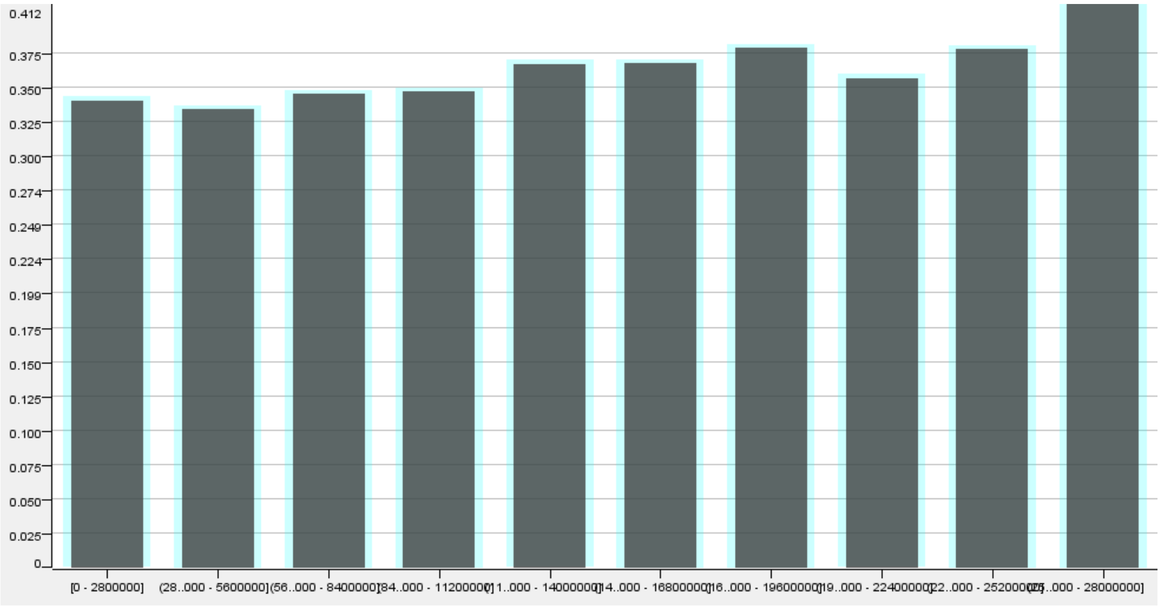


Figure 5: Average wOBA values for player salary ranges

The next step was to bin the data into salary ranges, and find the average offensive contributions for each bin. Salary bins were created using equal player distribution, rather than equal salary ranges. A crude method to determine player worth can be based on all players who posted above average statistics for their salary range and thus labeled as "worth their salary". But if a player is posting similar statistics to a player in a lower salary bracket, and they're both considered "worth it", then maybe more filters should be applied.

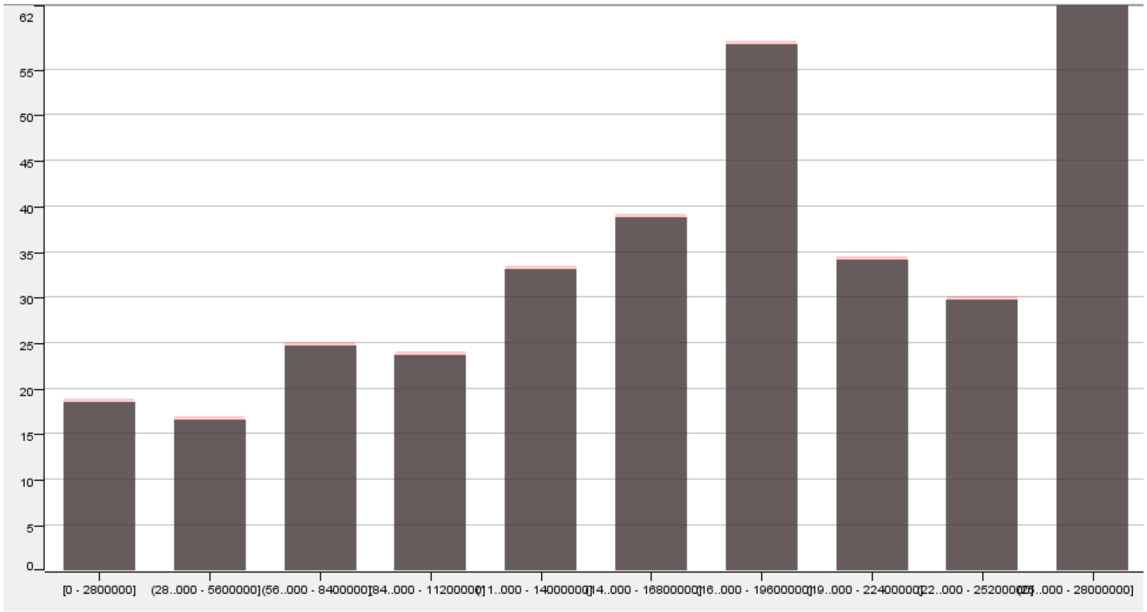


Figure 6: Average D_VORP for each salary bin

Weighted runs above average (wRAA) was the next useful statistic to be developed. It is derived from wOBA using the formula below:

$$\text{wRAA} = \left(\frac{(\text{wOBA} - \text{ league wOBA})}{\text{wOBA scale}}\right) \times \text{ Plate Appearances}$$

Where, league wOBA and wOBA scale are dependent on the MLB season. Based on research by Fangraphs, a baseball statistics website, a chart with player rating correlating to wRAA was developed [4].

Table 1: Per Fangraphs, wRAA vs. Player Rating

Rating	**wRAA**
Excellent	40
Great	20
Above Average	10
Average	0
Below Average	-5
Poor	-10
Awful	-20

A salary curve can be used, where there is a higher tolerance for players at lower salary ranges, and a lower tolerance for player production at higher ranges. If a player is above the average for his bracket, they still may be considered a poor contract.

In Figure 4 is the average player salary, is $4,349,423. At that point, if the wRAA is above 0, the player is worth their salary.

Experiment Setup

Using Excel the wOBA was calculated based on equation 1 and added as another column. wRAA was calculated using equation 2.

In KNIME the csv file was imported and only the most useful values were retained by using the column splitter node. Some values, RBI and OPS were normalized between 0 and 1 for visualization purposes. The scatter plot and histogram were the two main charts used for data visualization. Values such as wRAA and wOBA were used extensively for data visualization.

All rookies were removed from the dataset because they have different contract eligibilities than regular players and so their data may interfere with the predictors.

The first meta-node, conversions, contained a column splitter which removed all values that were not relevant, or had similar metrics already in the set. Salary was binned into three different bins for simplicities sake, then the original salary values were removed from the dataset.

A k-means clustering node was used to group players by wRAA and plotted on a scatter matrix to try and see a correlation between wRAA and salary.

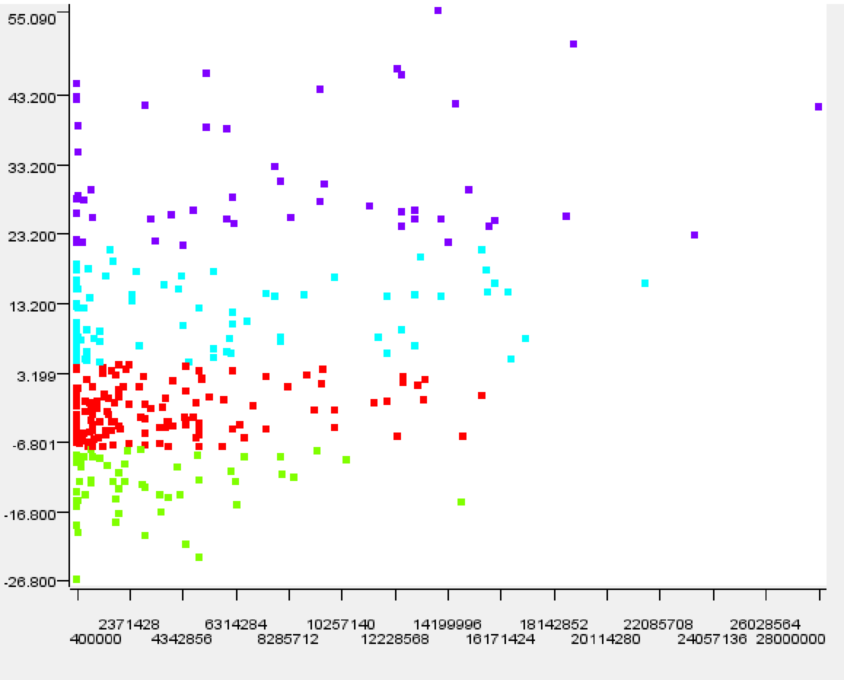
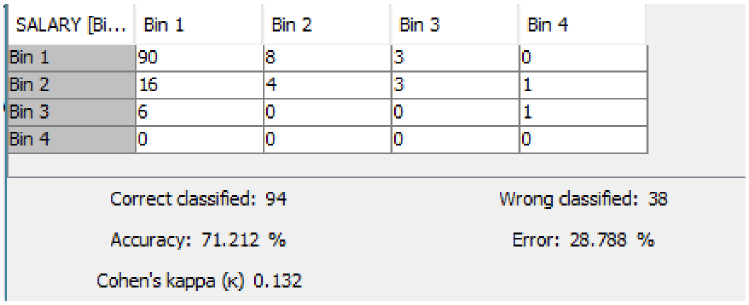


Figure 7: k-means clustering applied to wRAA and compared against player salary

From the graph (Figure 7) it can be seen that there is a slight trend, but there is still too much variance at the lower salary tiers.

The data fed into the learners was a combination of wRAA, RBI, wOBA and a normalized sum of RBI and OBP. The prediction columns targeted were the salary bins. The first run, four bins were created with equal width, and the predictors averaged around 70% accuracy. Classifications were good for the lowest salary bin, but as the salary increased results deteriorated (Figure 8).



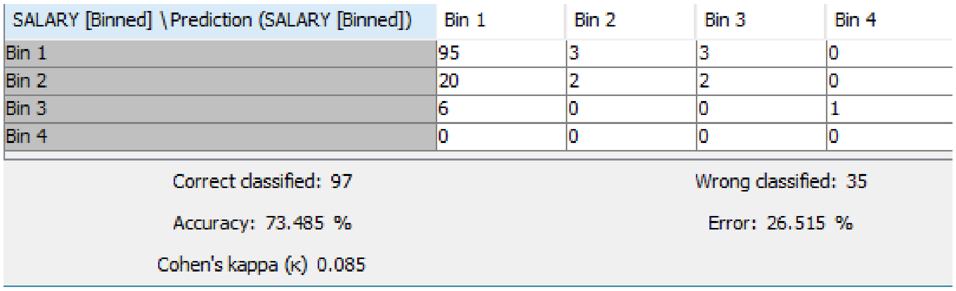
SALARY [Bi...	Bin 1	Bin 2	Bin 3	Bin 4
Bin 1	90	8	3	0
Bin 2	16	4	3	1
Bin 3	6	0	0	1
Bin 4	0	0	0	0

Correct classified: 94 Wrong classified: 38

Accuracy: 71.212 % Error: 28.788 %

Cohen's kappa (κ) 0.132

Figure 8: Confusion matrix for decision tree learner



SALARY [Binned] \ Prediction (SALARY [Binned])	Bin 1	Bin 2	Bin 3	Bin 4
Bin 1	95	3	3	0
Bin 2	20	2	2	0
Bin 3	6	0	0	1
Bin 4	0	0	0	0

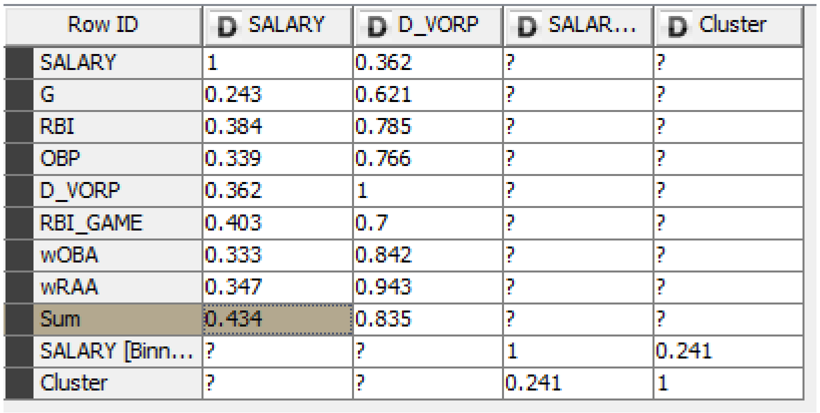
Correct classified: 97   Wrong classified: 35

Accuracy: 73.485 %   Error: 26.515 %

Cohen's kappa (κ) 0.085

Figure 9: Confusion matrix for the Random Forest operator

From the above figures (Figure 9), the decision tree learner and random forest learner were very similar in the output. A likely cause of this result is due to the distribution of players in each salary bin. Furthermore, the low correlation between statistics and salary influences the misclassifications, as evidenced by Figure 10.



Row ID	D SALARY	D D_VORP	D SALAR...	D Cluster
SALARY	1	0.362	?	?
G	0.243	0.621	?	?
RBI	0.384	0.785	?	?
OBP	0.339	0.766	?	?
D_VORP	0.362	1	?	?
RBI_GAME	0.403	0.7	?	?
wOBA	0.333	0.842	?	?
wRAA	0.347	0.943	?	?
Sum	0.434	0.835	?	?
SALARY [Binn...	?	?	1	0.241
Cluster	?	?	0.241	1

Figure 10: Show the correlation between statistics

Weak correlation values make it incredibly difficult to classify the data correctly. However, if defensive metrics were figured into the wRAA or WAR calculations, there would likely be a greater correlation value.

Predictors were run again on re-binned salaries of equal frequency and the results were much worse. Although, it is not surprising that there was a major decrease in the quality of the classification. Looking back at Figure 7, the k-means clustering, it is evident that issues will arise when classification occurs.

Figures 11, 12 and 13 in the appendix show the confusion matrix for bins of equal frequency, and the results are not very good. The best predictor managed 40% correct classification, which shows that there are serious flaws with the data quality.

Conclusions

The dataset was not complete, missing defensive statistics immediately removed half a player's worth from the equation. Therefore, even correctly classified salaries may not be considered deserving for the player if their defensive statistics drop their WAR from above average to below average. The predictors were able to accurately classify some players, but the results provided inconclusive evidence that it is possible to predict players' salary based on the given dataset.

References

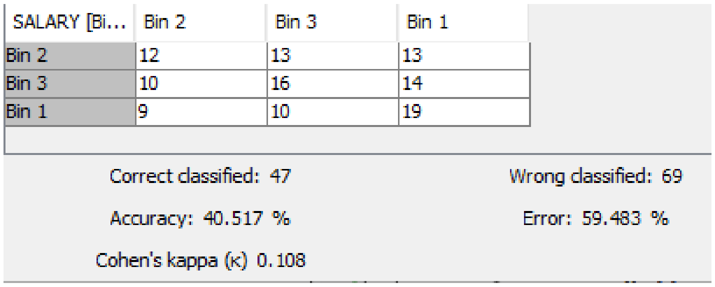
[1] S. Smalls, "How Long Until You're In the Bigs," 14 February 2011. [Online]. Available: http://minorleagueuniversity.blogspot.ca/2011/02/milb-life-how-long-until-youre-in-bigs.html. [Accessed 3 November 2015].

[2] D. Cameron, "How Fair is MLB's Pay Scale?," Fangraphs, 4 March 2013. [Online]. Available: www.fangraphs.com/blogs/how-fair-is-mlbs-salary-scale/. [Accessed 3 November 2015].

[3] T. Tango, A. Dolphin and M. Lichtman, The Book: Playing the Percentages in Baseball, Createspace Independent Pub, 2007.

[4] Fangraphs, "wRAA," Fangraphs, [Online]. Available: http://www.fangraphs.com/library/offense/wraa/. [Accessed 17 November 2015].

Appendix



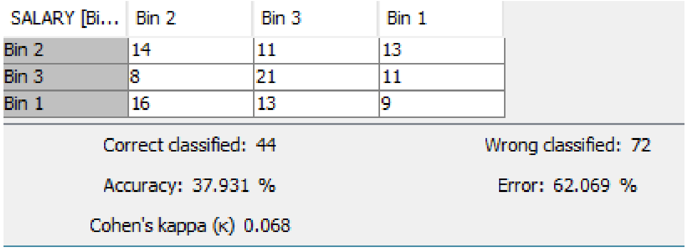
SALARY [Bi...	Bin 2	Bin 3	Bin 1
Bin 2	12	13	13
Bin 3	10	16	14
Bin 1	9	10	19

Correct classified: 47 Wrong classified: 69

Accuracy: 40.517 % Error: 59.483 %

Cohen's kappa (κ) 0.108

Figure 11: RandomForest results



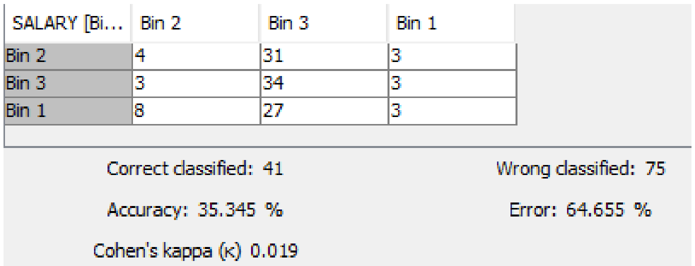
SALARY [Bi...	Bin 2	Bin 3	Bin 1
Bin 2	14	11	13
Bin 3	8	21	11
Bin 1	16	13	9

Correct classified: 44 Wrong classified: 72

Accuracy: 37.931 % Error: 62.069 %

Cohen's kappa (κ) 0.068

Figure 12: Decision Tree learner results



SALARY [Bi...	Bin 2	Bin 3	Bin 1
Bin 2	4	31	3
Bin 3	3	34	3
Bin 1	8	27	3

Correct classified: 41 Wrong classified: 75

Accuracy: 35.345 % Error: 64.655 %

Cohen's kappa (κ) 0.019

Figure 13: LibSVM results

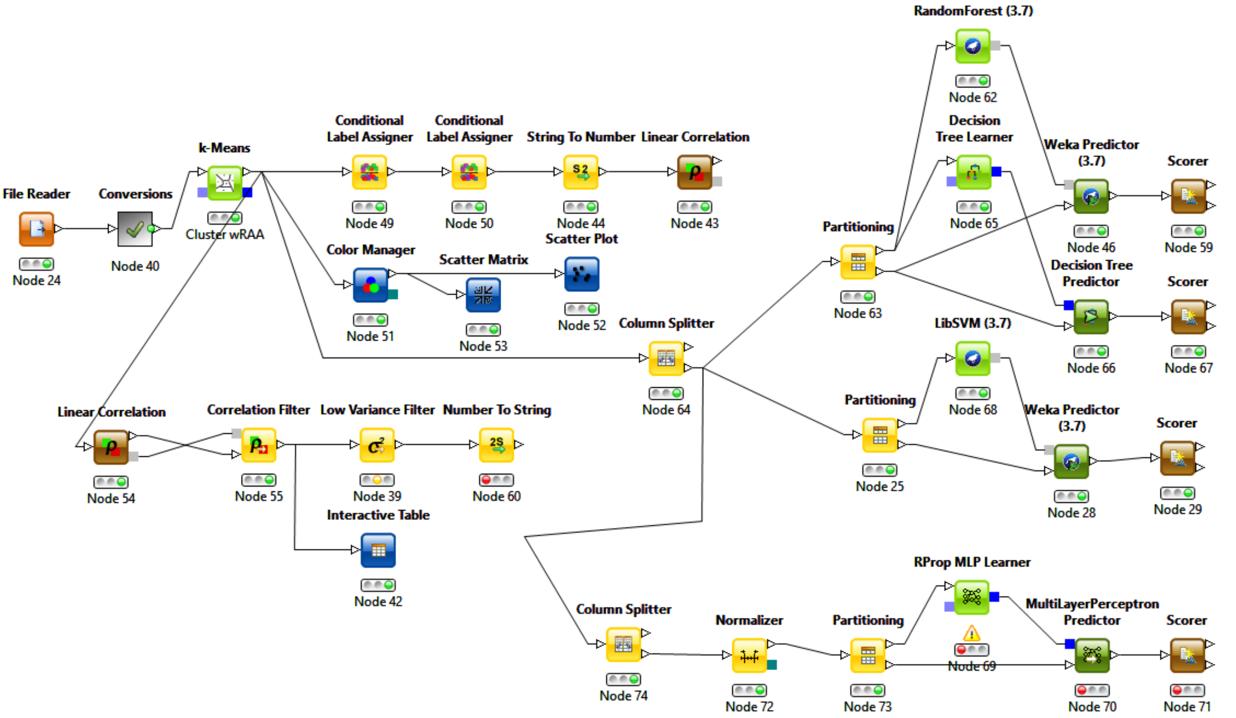


Figure 14: KNIME workflow visualization

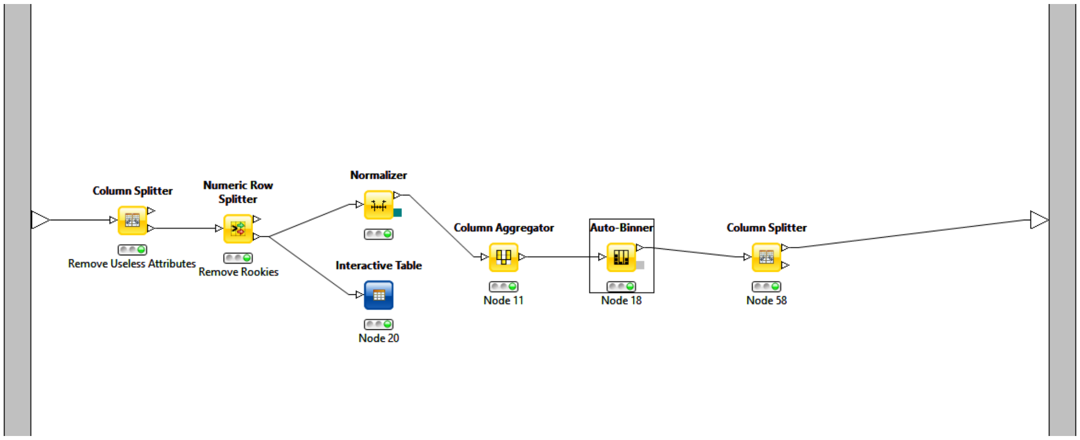


Figure 15: Conversion meta-node