An Approach to Predict Outcomes in Sports Games with Bigdata Techniques and Data Mash-Up

Jee-Ah Shin, Jin-Hwa Kim, Joo-Yong Lee

***Abstract*─The purpose of this study is to predict performance of players and teams in baseball games using data mining and data mash-up. In this paper, decision tree technique and data mash-up approach are used for predictions. A data set on 111 games by one of the most outstanding baseball player in S. Korea, player A(S.Y. Lee) and his team is collected for the study. Three sets of data are combined for the mash-up from 3 different sources: Korea Baseball Organization(KBO), Korea Meteorological Administration, and Google Trends. The results from the analysis have 3 findings. Firstly, the important variables for 'H(Hits)' are google trends and humidity. If 'trend' is 25 and more and 'humid' is 77.1 or over, the probability to make one or more hits is 85.7%. It is most likely for player A to make one or more hits when public interest is high and the weather is humid. Secondly, the number of spectators and humidity are significant for 'BB(base on balls)'. If 'spect' is 12823.5 or more, 'humd' is 62.2 or above and 'humd' is below 73.2, probability to get 'BB' is 57.1%. When there are many spectators and it is moderately humid, the probability for getting 'base on balls(walk)' is high. Thirdly, wind speed and temperature are important to have a good 'result'. If 'wind' is 2.75 or over and 'temp' is 25.1 or above, probability to get a team victory is 100%. When the wind blows a little and the temperature is high, his team will win. This study focuses on a baseball team and a player. Further study can extend the scope of applications to other teams and other sports*.**

***Index Terms*─Sports game, prediction, data mining, data mash-up, decision tree**

I. INTRODUCTION

Big data became one of the mainstream techniques in these days and it is a popular analytical method in baseball games not only in S. Korea but also in many other countries including US. In the US, as a leader this trend, ZiPS (sZymborski Projection System) is a player estimation system developed by Dan Szymborski of ESPN. It was created when he was at Baseball Think Factory and can be found on the Baseball Think Factory website. ZiPS makes growth and decline analysis curves for each player. The system uses statistical data from the previous four seasons for 24 to 38 year old players. And it analyzes the data of recent games with more weight. The system uses statistics from the previous three years only for younger or older players. This uses also the data on velocities of balls, injuries, and play-by-play data[1]. In summary, ZiPS uses past performance and aging trends to predict players' future performance and it is used to predict the performance of players in the remaining seasons, which is updated daily. Previously, the vast majority of people did not have much interest in ZiPS's forecasts, and did not expect the prediction would be accurate. However, ZiPS is now recognized as one of the most accurate prediction systems. Also in S. Korea, the Korea Baseball Organization league is gradually increasing its use of statistical analysis. Big data on player records also is stored in the KBO league database. But the level of technologies and applications are low because the history of KBO is short compared to MLB, and the number of teams and players are small. However, the interest in prediction of a baseball game is increasing over time[2].

Baseball is one of the most popular sports in all ages and genders all over the world. The history of Korean baseball is over one hundred years since 1905. Especially after the beginning of the Professional Baseball in 1982, it became more popular. As batting and catch a ball alone with basic motions like a run, a jump, and a throw, it quickly became a national sports game in S. Korea. And partnership, judgment, and determination have an effect on the outcome of a game. People want their team to win the game and also want their favorite players to show excellent plays in each game. In S. Korea, many people like *player A(S.Y. Lee)*, the No.1 hitter among so many great baseball players. He is a very popular as a representative cleanup hitter, the king of home run, and also a plodding player. But *player A* didn't get good results all the time. The outcome of a baseball game is influenced by many factors such as conditions of players and the external factors such as weather, the number of spectators, and public opinions. This study suggests a model that is used to predict outputs in each game and also to predict whether *player A* gets a good score or not under certain circumstances. This study has two objectives:

- To select the external factors that can have an effect on a baseball game and look into over correlation between the factors and scores using decision tree technique.
- To improve accuracy of prediction on hits, base on balls(walk) and game outputs by *player A* based on the result of analysis.

Revised Manuscript Received on May 22, 2019.

Jee-Ah Shin, School of Business, Sogang University, 1 Shinsoo-Dong, Mapo-Gu, Seoul, Korea,

Jin-Hwa Kim, School of Business, Sogang University, 1 Shinsoo-Dong, Mapo-Gu, Seoul, Korea,

Joo-Yong Lee, School of Business, Sogang University, 1 Shinsoo-Dong, Mapo-Gu, Seoul, Korea,

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

Among the data mining techniques, the classification technique is particularly useful when creating rules for prediction of outcomes[3]. This study uses decision tree technique that is an easy and practical classification method to build a significant prediction model that predicts result of a game under various factors, especially external factors[4]. Decision tree is represented in the form of a flowchart with the tree structure. Each node explains a test about an attribute. Leaves mean class labels and branches stand for conjunctions of characteristics that lead to each class label. Decision tree is simple in extracting each classification rule. Decision tree describes rules for dividing data into groups. It is considered to be an accurate method in comparison with other classification methods. A number of data mining techniques have already been used to improve the performance such as regression, k-means clustering, associate rules, and prediction methods[5], [6]. Data mining techniques are used in a various fields to identify factors, extract variables, create rules, and improve people's understanding of customer behavior[7]. Compared to general statistical techniques, decision tree method has many advantages in classifying and predicting outputs according to rules from trees. In other words, it is easy to apply model to prediction compared to other methods such as regression analysis and discriminant analysis. The model from decision tree can be used immediately without conversion of independent variables. Despite the above advantages, there is instability that the tree structure can be changed even with the small change of the data. But there is no big difference in the accuracy of the whole tree according to the change in the lower level. Also, the predicted value can be greatly changed with a small change around the threshold of the main classification variable because it has a difficult to split around the classification value[8].

When we want to combine data or information with other data or information, mash-up is used. Mash-up is a technique, websites or web applications where they use data, presentation or functionality from two or more sources to create a new service by using them[9]. It is possible because web services or public APIs allow free access. It is a term in music that is used to create a new song from mixing two or more songs. In the field of information technology, it means a development of new services by combining various information, contents, and services provided by web service providers in the web[10]. In other words, it stands for creating contents and services of a new dimension with contents of different websites. It is possible to develop a new application service or mash-up by blending a unique user interface or contents based on APIs disclosed by web service providers. Because many major companies have released APIs that allow their content to be used, many mash-up services have been created. The advantage of mash-up is that the cost of administering new services is very low as they make use of already existing resources. And it is useful to developers as it doesn't require much code. A major weakness is that it is dependent on other services, and mash-up is also stopped when services of primary resources are disrupted. There is a management difficulty because it is necessary to change the form of provision of the primary resource when it changes its service[11]. In the age of Web 3.0, the concept of mash-up is getting more and more popular. Data mash-up means a mix of various data from different data sources. It is a combination of different data from different source in different types, which uses its own open application programming interfaces. It is sometimes necessary to combine different data sets to achieve a business goal.

Previous studies related to a baseball prediction are as follows. In Tae Young Yang and Tim Swartz[12]'s paper, A Markov Chain and a Monte Carlo algorithm were used in implementing Bayesian inference and simulating results of upcoming games. Winning in the major league baseball is decided by various factors and they are associated with the past outcome such as the batting ability of the two teams and the starting pitchers. The goal of Arlo Lyle[13]'s research is to find out how players' results change over time. Based on the results from this analysis, it is used to predict performance of the players. For the prediction, past 30 years statistics of players have been used. The result in this paper shows that machine learning techniques are similar in their accuracy to the best predictive models in predicting players' performance. Wenhua Jiang and Cun-Hui Zhang[14] use empirical Bayes methods for prediction of batting averages with 2005 Major League baseball data. Empirical Bayes estimators are recent and general in homoscedastic and heteroscedastic partial linear models. Chenjie Cao[15]'s project focuses on using machine learning algorithms to build models for predicting outcomes from the NBA games and theses algorithms includes Simple Logistics Classifier, Artificial Neural Networks, SVM, and Naïve Bayes. In order to produce a convincing result, a data set from 5 regular NBA seasons is collected for training data and data from 1 NBA regular season is used as test dataset. Randy Jia, Chris Wong, and David Zeng[16] want to predict baseball games with higher accuracy. And they try to find information on which factors lead the baseball game team to a victory. This study selects characteristics related to baseball games and analyzes the data using classification and regression techniques. Brandon Tolbert and Theodore Trafalis[17] build a model using data mining methods. Their model shows better performance in predicting World Series winners, American League champions, and National League champions compared to other traditional models. This paper uses data from regular season with kernel machine learning schemes. They also try to find useful features that can help forecast future winners.

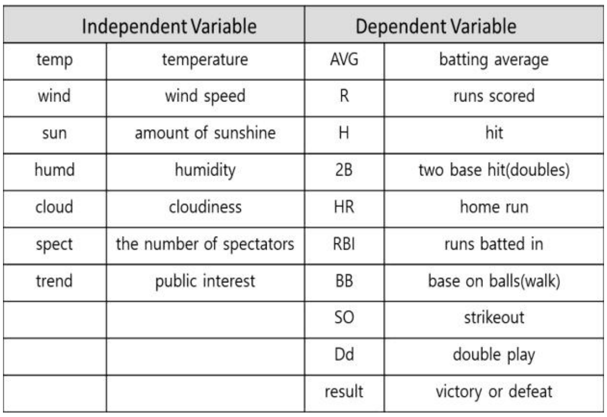
II. DATA COLLECTION AND METHODS

A. Data collection

The suggested model in this paper has independent variables and 10 dependent variables as shown in Table 1.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table. 1 Variables



Independent Variable		Dependent Variable	
temp	temperature	AVG	batting average
wind	wind speed	R	runs scored
sun	amount of sunshine	H	hit
humd	humidity	2B	two base hit(doubles)
cloud	cloudiness	HR	home run
spect	the number of spectators	RBI	runs batted in
trend	public interest	BB	base on balls(walk)
		SO	strikeout
		Dd	double play
		result	victory or defeat

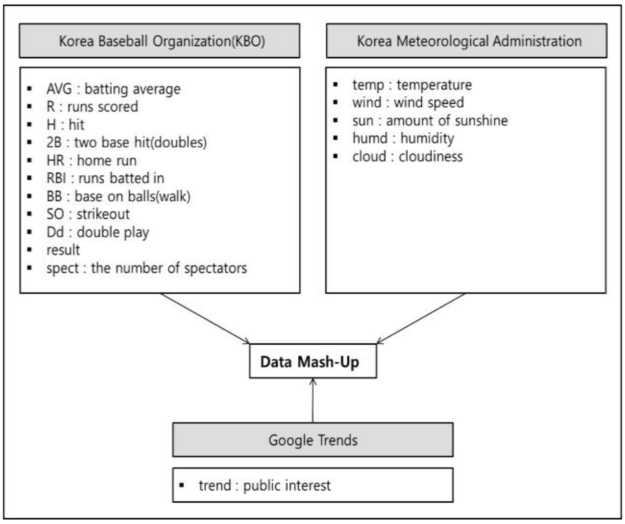


Fig. 1 Data Mash-Up

Figure 1 shows mash-up of data from various sources like Korea Baseball Organization(KBO), Korea Meteorological Administration, and Google Trends. Total 111 instances contain information on score from KBO, weather from Korea Meteorological Administration, and public opinion from Google Trends. The information on batting average, runs scored, hit, two base hit(doubles), home run, runs batted in, base on balls(walk), strikeout, double play, and result for *player A* is collected and saved. The information on weather includes temperature, wind speed, amount of sunshine, humidity, and cloudiness. The number of spectators for each game is also collected.

B. Methods

As you see in Figure 2, 10 decision trees are produced for each of 10 dependent variables as a target variable. Among the these decision trees, we came up with the 3 significant outcomes: ‘H(hit)’, ‘BB(base on balls)’, and ‘result’. The overall research procedure in this study is shown in Figure 3. Firstly, data related to research is collected. Secondly, data sets from 3 different sources are merged into one data set, which is called data mash-up. Thirdly, decision tree analysis is performed for each dependent variable. Fourthly, the decision trees with ‘hit’, ‘base on balls’, and ‘result’ as a target are explained in detail for further applications. Finally, the important rules in each of the three dependent variables are identified.

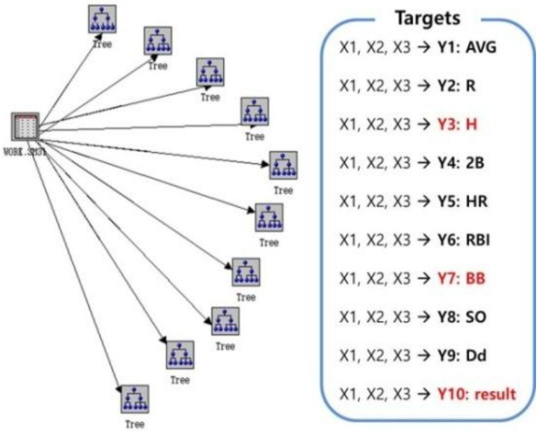


Fig. 2 Decision Tree Model

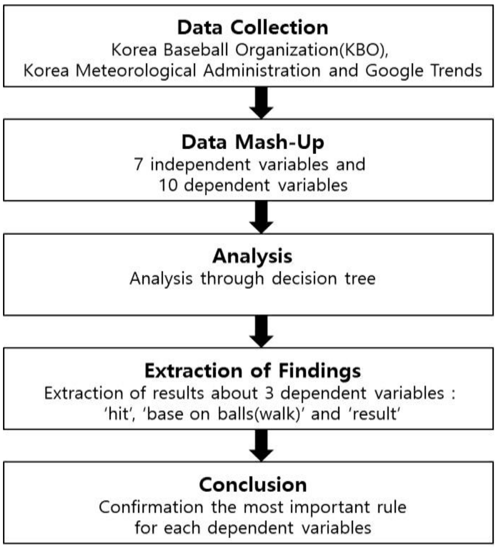

Data Collection
Korea Baseball Organization(KBO),
Korea Meteorological Administration and Google Trends
Data Mash-Up
7 independent variables and
10 dependent variables
Analysis
Analysis through decision tree
Extraction of Findings
Extraction of results about 3 dependent variables :
‘hit’, ‘base on balls(walk)’ and ‘result’
Conclusion
Confirmation the most important rule
for each dependent variables

Fig. 3 Research Process

III. RESEARCH ANALYSIS AND RESULTS

A. A decision tree with ‘H(hit)’ as a target

Figure 4 shows the result of analysis on a decision tree with ‘H(hit)’ as a target. The meaning of the each number from 0 to 3 inside the square in decision tree is shown in the Table 2. There are six rules in the graph. Firstly, the important variable deciding ‘H’ is ‘trend’. When ‘trend’ is less than 25, the probability to make one or more hits is about 65.6%. When it is 25 and more, the probability to make one or more hits is 60%. Secondly, relevant factors are ‘temp’ and ‘humd’. When ‘temp’ is 23.55°C or above, the probability to make one or more hits is 81.8% and when ‘humd’ is 77.1 or over, it is 85.7%. Thirdly, other significant variables, alone with ‘trend’ and ‘humd’, are ‘cloud’ and ‘temp’. When ‘cloud’ is below 2.05, the probability to make one or more hits is 66.7%. When ‘temp’ is under 31.45°C, the probability to make one or more hits is 82.1%. As a result, the most significant rule among them is as follows. We can finally conclude the condition for *player A*’s best play on hit. If ‘trend’ is 25 or more and ‘humd’ is 77.1 or over, then the probability to make one or more hits is 85.7%. It is most likely for *player A* to make one or more hits when

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

public interest is high and the weather is humid.

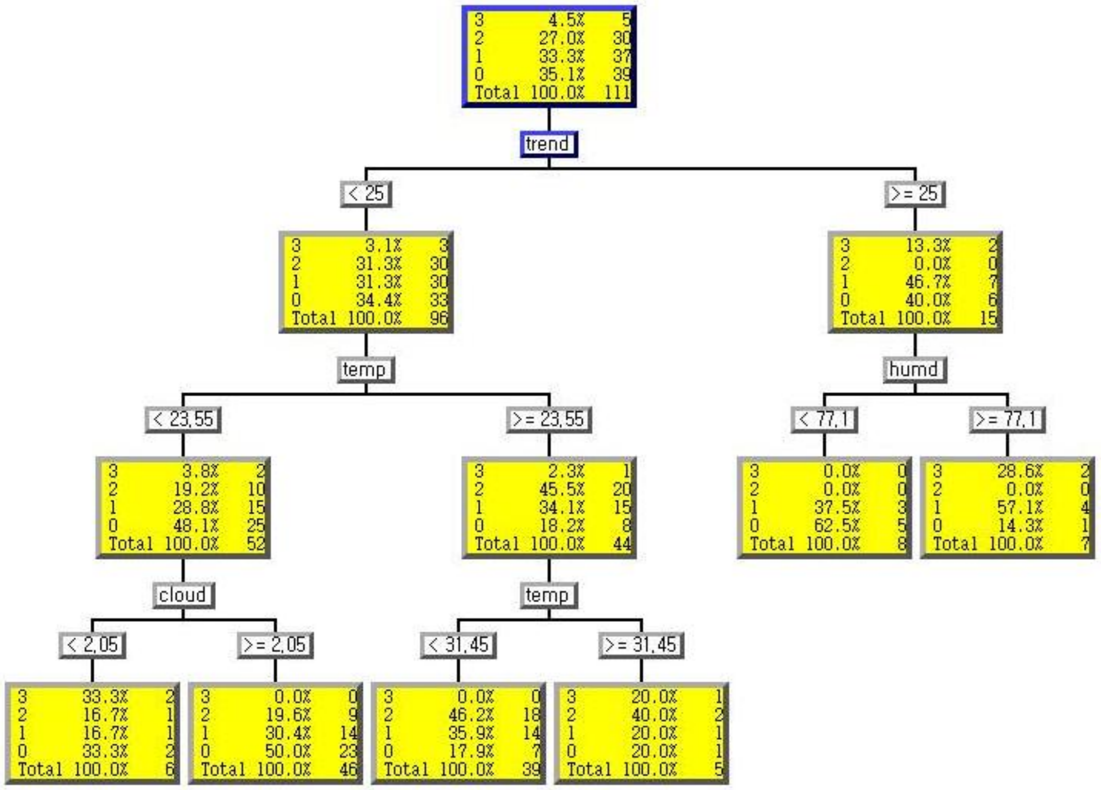
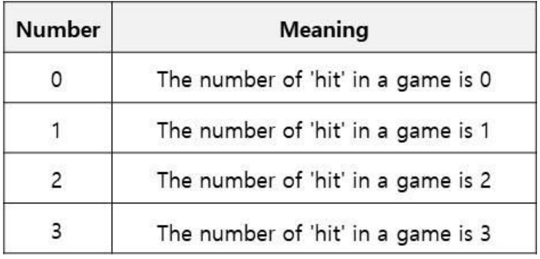


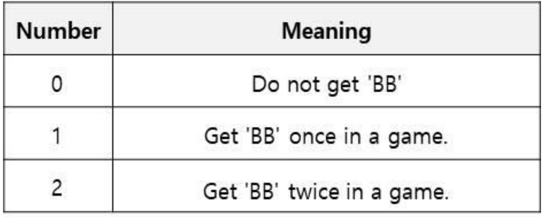
Fig. 4 A decision tree with 'H' as a target

Table. 2 Meaning of number for each value of 'H'



Number	Meaning
0	The number of 'hit' in a game is 0
1	The number of 'hit' in a game is 1
2	The number of 'hit' in a game is 2
3	The number of 'hit' in a game is 3

Table. 3 Meaning of number for each value of 'BB'



Number	Meaning
0	Do not get 'BB'
1	Get 'BB' once in a game.
2	Get 'BB' twice in a game.

B. A decision tree with 'BB(base on balls)' as a target

Figure 5 shows a decision tree with 'BB(base on balls)'. The number 0, 1 or 2 in the square represent the number of 'BB' in the game as seen in Table 3. 0 means not getting 'BB', 1 and 2 mean the number of 'BB' by *player A.*

Number of rules in the figure is 7. Firstly, the most important variable deciding 'BB' is 'spect', the number of spectators. When 'spect' is 12823.5 or more, the probability to get 'BB' is 40.9%. Secondly, 'humd' is next important decision variable. When 'humd' is 62.2 or above, the probability to get 'BB' is 50%. Thirdly, other important variables deciding 'BB' are 'spect' and 'cloud' alone with 'spect' and 'humd'. When 'spect' is 9860 or more, the probability to get 'BB' is 28.6%. When 'cloud' is less than 9.5, it is 32%. Finally, when 'humd' is below 73.2, the probability to get 'BB' is 57.1%. The most meaningful rule is that if 'spect' is 12823.5 or more, 'humd' is 62.2 or above and 'humd' is below 73.2, then probability to get 'BB' is 57.1%. The probability of getting 'BB' is high in crowded and humid weather for *player A.*

C. A decision tree with game 'result' as a target

The highlight of a baseball game is game 'result', which is victory vs. defeat. Figure 6 shows a decision tree that explains factors influencing the result of each game. As in Table 4, number 1 means a defeat, 2 means a victory, and 3 means a tie in the game. Firstly, the study finds that the most important variable deciding game result is 'wind'. When 'wind' is 2.75 or over, the probability to win is 79.4%. Secondly, other important variables deciding game results are 'cloud' and 'temp'. When 'cloud' is 2.85 or more, the probability to win a victory is 50% and when 'temp' is 25.1 or above, the probability to win is 100%. Thirdly, 'spect' and 'wind' are also important variables deciding game results. When 'spect' is 8506 or more, the probability to win is 91.7% and when 'wind' is less than 1.45, the probability to lose is 81.8%. Among seven rules, the best rule for winning the game is that if 'wind' is 2.75 or over and 'temp' is 25.1 or above, then probability to win is 100%. When the wind blows a little and the temperature is high, the *player A* has best chance to win the game.

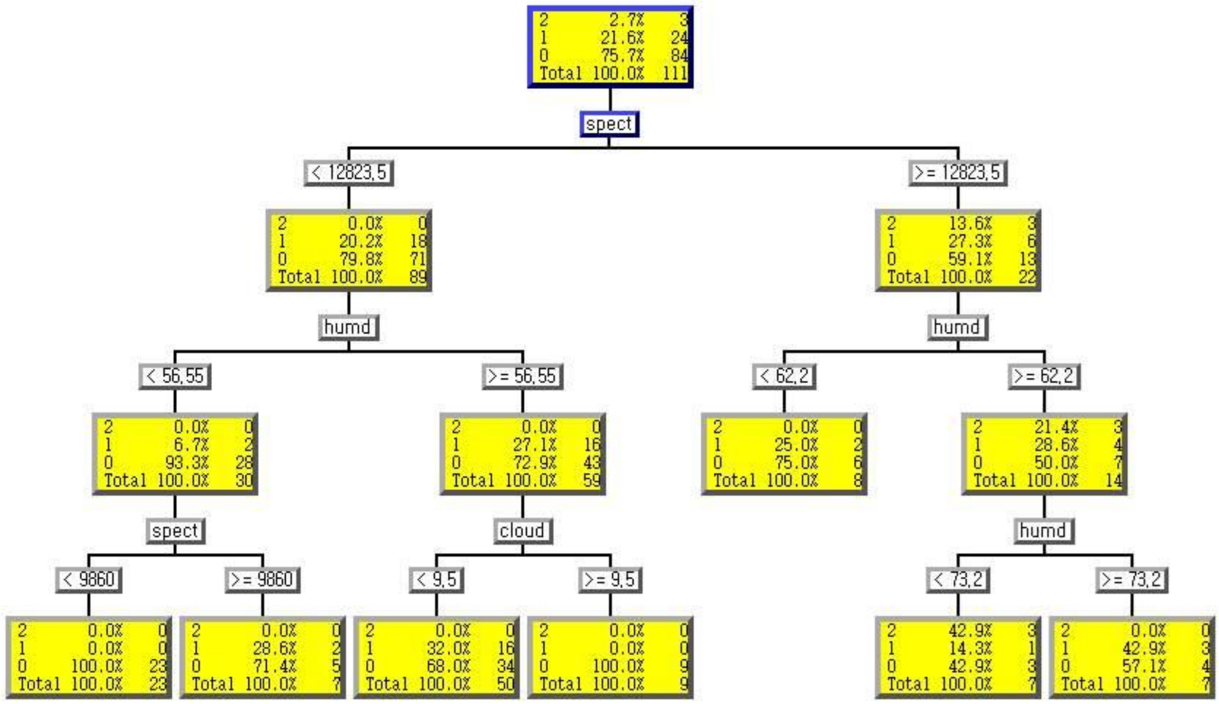


Fig. 5 A decision tree with 'BB' as a target

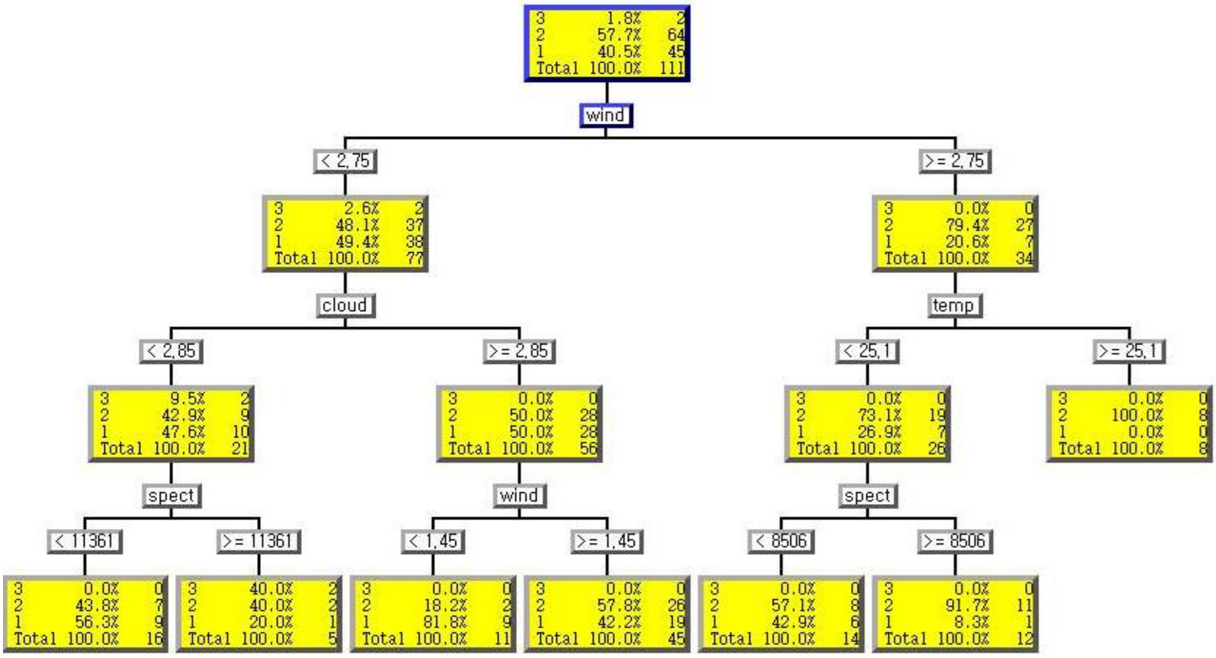


Fig. 6 A decision tree with game 'result' as a target

Table. 4 Meaning of number for each value of 'result'

Number	Meaning
1	a defeat
2	a victory
3	a tie

IV. CONCLUSION

This study uses decision tree analysis to predict game outputs with variables influencing baseball games. Especially, the study focuses on *player A*, the representative baseball player in S. Korea. There are 7 independent variables that are related to the external factors: temperature, wind speed, amount of sunshine, humidity, cloudiness, the number of spectators, and google trends. These data sets are collected from 3 data sources and these data sets are consolidated into one data set as data mash-up. The tests show that decision tree models can predicts 'hit', 'base on balls', and 'result' with reliability. The results from this study are shown next.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Firstly, the outputs from the analysis show that google trends, temperature, humidity, and cloudiness are important variables regarding game results from a decision tree with ‘H’ as a target. The best conditions for ‘H’ are high public interest and quite humid weather. In this environment, the probability of making more than one hit is 85.7%.

Secondly, a decision tree with ‘BB’ as a target shows that 3 variables are found to be important variables. *Player A* needs to have a specific condition in order to get ‘BB’. There should be a large number of spectators who come to watch a baseball game and the humidity must be slightly high. When these conditions are satisfied, the probability of getting ‘BB’ more than once is 57.1%.

Thirdly, a decision tree with ‘result’ as a target shows conditions in which the *player A* win the game with high probability. The variables influencing the game ‘result’ are wind speed, cloudiness, temperature, and the number of spectators. In order for *player A* to win the game, the wind must blow and the temperature should be a little warmer. When these conditions are satisfied, the probability of winning the game is 100%. This means the chance of winning the game is highest in this condition.

The results from this research show that a data mining technique such as a decision tree can be used to predict the outcomes of the baseball games. The study also shows that combining different data sets from different data sources, which is called data mash-up, can be applied to predictions in diverse areas such as a baseball games.

The limitation of this paper is that we use only external variables with the assumption that the internal variables are not significant in prediction of game results. In addition, there may be other variables that can influence the outcomes of the games. Also the data used in this paper is only from the game in which *player A* played. The scope of data can be extended to more players with additional decision variables influencing the outcomes of games.

Data mining and data mash-up can be applied to more complicated prediction problems. As interests in predicting sports games are growing, the suggested methodology in this study can be applied to other sports games.

REFERENCES

1. “MLB - STATS » Glossary » Projection Systems » sZymborski Projection System(ZiPS).”[Online]. Available: http://m.mlb.com/glossary/projection-systems/szymborski-projection-system. [Accessed: 01-Apr- 2018].
2. “KBO – TEAMS » Player Search » Player Info.” [Online]. Available: http://eng.koreabaseball.com/Teams/PlayerSearch.aspx. [Accessed: 01-Apr- 2018].
3. D. Reby, S. Lek, I. Dimopoulos, J. Joachim, J. Lauga, and S. Aulagnier, “Artificial neural networks as a classification method in the behavioral sciences,” Behavioral Processes, vol. 40, no. 1, pp. 35-43, 1997.
4. M. Haghighat, H. Rastegari, and N. Nourafza, “A Review of Data Mining Techniques for Result Prediction in Sports,” ACSIJ Advances in Computer Science: an International Journal, vol. 2, no. 5, pp. 7-12, 2013.
5. J. R. Quinlan, “Induction of Decision Trees,” Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.
6. S. Baik and J. Bala, “A Decision Tree Algorithm for Distributed Data Mining: Towards Network Intrusion Detection,” International Conference on Computational Science and Its Applications, pp. 206-212, 2004.
7. C. K. Leung and K. W. Joseph, “Sports data mining: predicting results for the college football games,” Procedia Computer Science, vol. 35, pp. 710-719, 2014.
8. H. Sharma and S. Kumar, “A Survey on Decision Tree Algorithms of Classification in Data Mining,” International Journal of Science and Research(IJSR), vol. 5, no. 4, pp. 2094-2097, 2016.
9. D. Guinard and V. Trifa, “Towards the Web of Things: Web Mash-Ups for Embedded Devices,” In Workshop on Mash-Ups, Enterprise Mash-ups and Lightweight Composition on the Web(MEM 2009) in proceedings of WWW(International World Wide Web Conferences), vol. 15, 2009.
10. J. Yu, B. Benatallah, F. Casati, and F Daniel, “Understanding Mash-up Development,” IEEE Internet Computing, vol. 12, no. 5, pp. 44-52, 2008.
11. G. D. Lorenzo, H. Hacid, H. Y. Paik, and B. Benatallah, “Data Integration in Mash-Ups,” SIGMOD Record, vol. 38, no. 1, pp. 59-66, 2009.
12. T. Y. Yang and T Swartz, “A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball,” Journal of Data Science, vol. 2, no. 1, pp. 61-73, 2004.
13. A Lyle, “Baseball Prediction Using Ensemble Learning,” University of Georgia, 2007.
14. W. Jiang and C. H. Zhang, “Empirical Bayes in-season prediction of baseball batting averages,” In Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown. Institute of Mathematical Statistics, vol. 6, pp. 263-73, 2010.
15. C. Cao, “Sports Data Mining Technology Used in Basketball Outcome Prediction,” Dublin Institute of Technology, 2012.
16. R. Jia, C. Wong, and D. Zeng, “Predicting the Major League Baseball Season,” CS229 FINAL PROJECT, 2013.
17. B. Tolbert, and T. Trafalis, “Predicting Major League Baseball Championship Winners through Data Mining,” Athens Journal of Sports, vol. 3, no. 4, pp. 239-52, 2016.