Paper ###-2019

**The (‘Data’) Science of Chess – A Study of Factors Influencing the Game’s Outcome**

Alakananda Giridhar, Himani Reddy, Rahi Sheth, Prof. Kam Tin Seong

Singapore Management University

# Abstract

Chess has been one of the most invested areas of AI in the past, since it is primarily a game of patterns, and ample information is contained within each game. Therefore, this study has been dedicated to understand the patterns in a game of chess and assess the techniques and moves leading a (white) player towards winning or losing, using various data analytics techniques. The data collected provides data on several key variables influencing the victory status like the player rating, moves, number of turns and opening moves, among many others. Using exploratory analysis insights such as the most frequently used move to checkmate the black king, the most used opening move that gives the highest probability of winning and opening move preferences of proficient players etc were derived. Associative analysis identified moves with high support and move pairs with considerable confidence and lift, and Sequence Discovery concurred with it by observing the winning sequence of moves. Finally, multinomial and binary logistic regression enabled deduction of the impact of a certain move’s frequency on the outcome of the game determined by the odds ratios.

# 1. Introduction

Chess is a game of intellect, strategy, and one of patterns and sequences. Capitalizing on this, we analyzed the moves in every game, in order to list the dos and don’ts of the game, winning and losing move sequences etc. The techniques used to achieve them are Associative Analysis, Sequence Discovery and Multinomial Logistic Regression, the results of which are discussed in further sections.

# 2. DATA SOURCE

The dataset has been acquired from Kaggle which consisted of data regarding over 20K games extracted from an online chess portal called lichess.org

# 3. approach



# 4. DATA PREPARATION

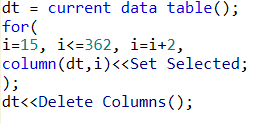
## 4.1 DATA CLEANING

The data acquired from Kaggle presents it in a format that is easy to process, for example, what allows a player to win as black or white and what factors affect a game (Number of turns, types of step patterns, opening phase, time duration, opening move type etc.). Each record had the moves of both the players’, from a particular game.

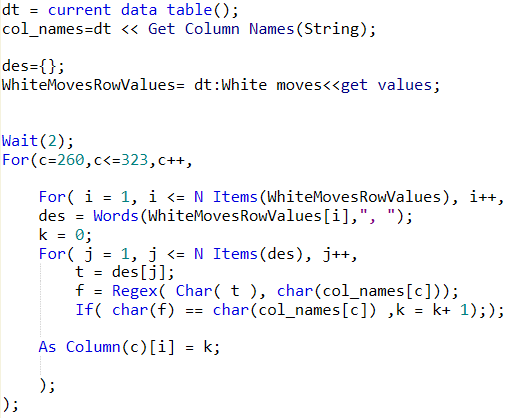
Hence, for this analysis, the moves of white players' alone have been extracted by delimiting and excluding every ***even*** move, since they correspond to those of black players.

The following steps have been performed using the **SAS JMP Pro 14.0.0 Toolkit.**

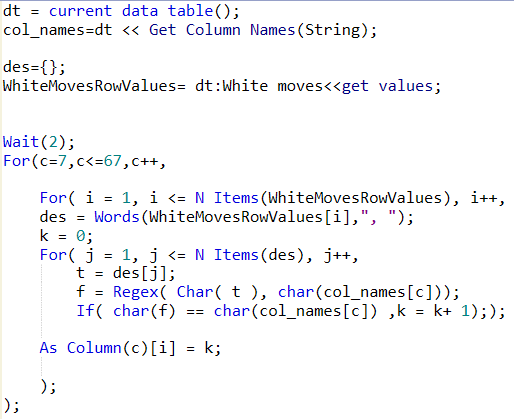
* Split “moves” column into individual columns using **Text to Columns** feature
* Retain white’s moves alone using the following jsl script:



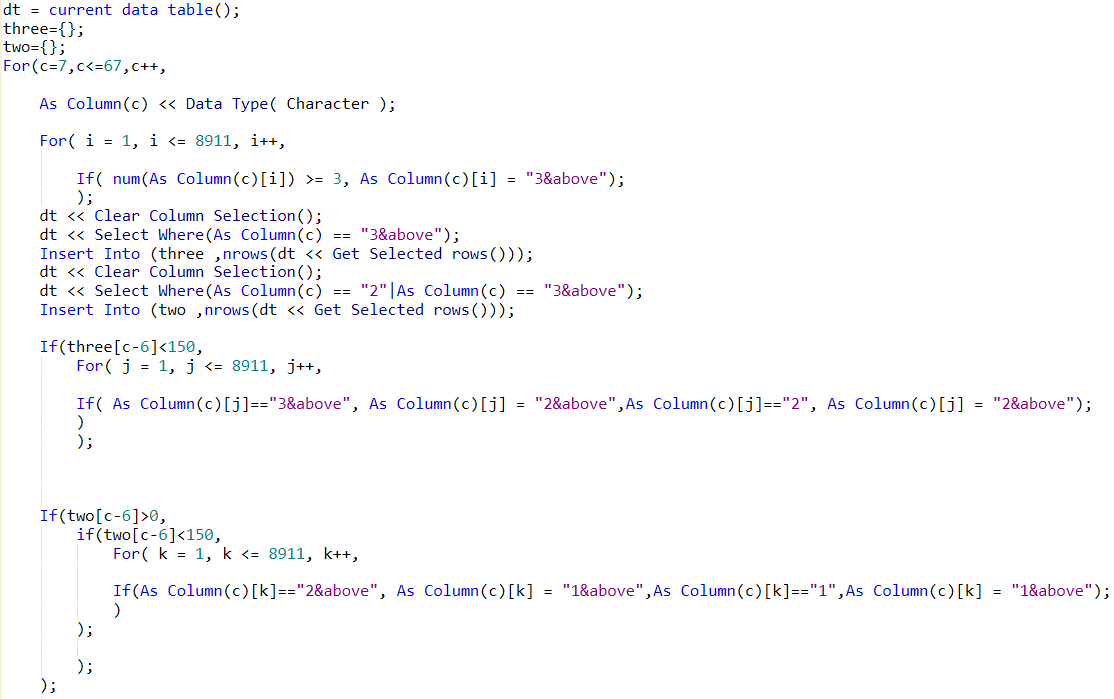
* Combine the columns using **Combine Columns** feature
* To visualize the chessboard utilization of win/loss/draw games, the frequency of squares **a1** to **h8** for each game/row was calculated from the derived “White moves” column using the following jsl script:



* The most popular moves were identified by observing the distribution of “White Moves”, and a column was created for each of them
* The frequencies of these popular moves (columns) within each game was calculated using the following jsl script:



* Recoding - Since most of the moves are frequented either 0,1 or 2 times in each game, frequencies of 3 and above were binned into category, “3&above”. If the count of games with frequency “3&above” was <150, it was recombined into a category called “2&above”. The same logic was used to create another category “1&above” using below script.



* There were no missing values in the data

## 4.2. DATA EXTRACTION

* It was observed that few records in the data have **Victory Status** as "**Resign**".
* A further study explained that these records describe games with no legal/fair winner, i.e., the games where a player abruptly withdraws from the game, steering the opponent player to win.
* Hence, such games have been excluded from the analysis for unbiased results.

# 5. Descriptive statistics

Chess is an interesting game as it is, an intense study has been done to analyze the factors influencing the game as well the prominent features of the game here.

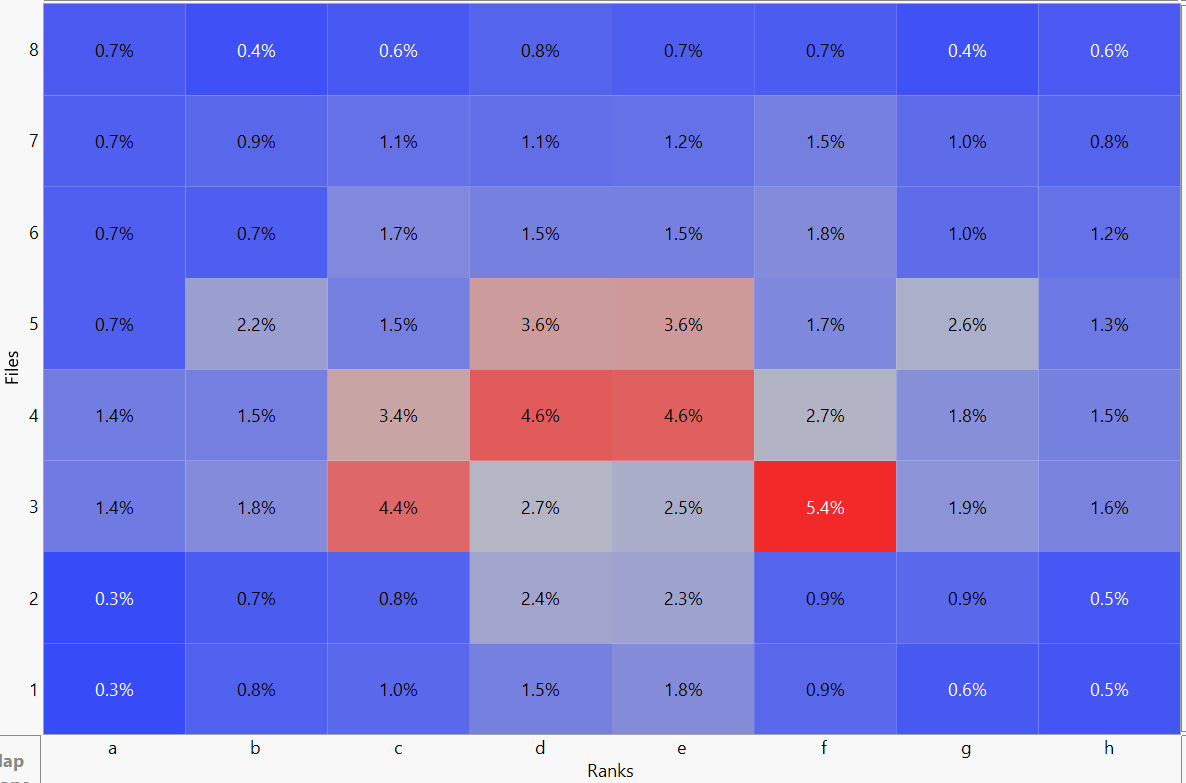


Figure 2. Board distribution of white’s moves from their winning games

Above heatmap depicts that the center of the board is prudently used pressing on the fact that occupying or advancing more into the center of the chess board translates to having better odds in the game. This scenario captures the white’s effective utilization of first mover’s advantage. The opposite pattern has been observed in the losing matches of white.

The move which has been frequently used to check mate by white players has been identified as **Qxf7#** which is also called as **Parham Attack or Scholar's Mate**. The Black King cannot take the White queen because the White Bishop on c4 defends it (Refer Appendix for chessboard layout). Further, moving the Black King to e7(Ke7) does not allow it to escape capture on the next move.

The most used opening moves were found as **Sicilian Defense** where the pawn is moved to e4 and has been the most recommended chess opening as it occupies the center position giving an advantage for the white player and when used properly is observed to have higher weightage in winning the game. The other prominent chess openings are found as *French Defense, Queen’s Pawn game* and *King’s Pawn game*

The proficient players (higher rating) were observed to have used not so frequent opening moves such as **Dutch Defense** where d4 move is played first followed by f5 from Black’s side stakes a serious claim to the e4 square and envisions an attack in the [middlegame](https://en.wikipedia.org/wiki/Chess_middlegame) on White's [kingside](https://en.wikipedia.org/wiki/Glossary_of_chess#kingside); however, it also weakens Black's kingside some and contributes nothing to Black's [development](https://en.wikipedia.org/wiki/Glossary_of_chess#development).

Majority of games are observed to have taken 50-60 turns throughout the game’s progress as observed by the normal distribution of the moves from all the games.

The figure below(Figure 3) gives the most used closing move to checkmate the black king.

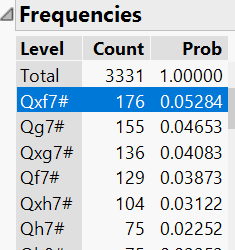


Figure 3. Frequently used closing move

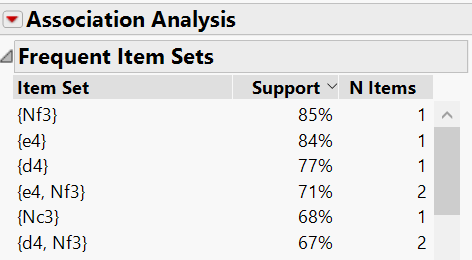
# 6. METHODOLOGY

## 6.1 ASSOCIATION ANALYSIS

Association analysis is done to explore the associations among moves to identify patterns in player’s game behavior. In general, many amateurs memorize the first few most popular opening moves in chess, but It’s important to understand that whenever a player makes a move, they are gaining or losing control over certain critical areas of the board.

Drawing one step closer to the motive, we adhered to associative analysis in JMP to explore the associations among moves and their prominence.

The moves with high support are found to be e4, d4, Nf3 and followed by castling move (O-O). Interestingly amongst the individual opening moves of all the win/loss and draw games it has been found that the moves like g3 and f4, d3 and b3 are observed with more frequency in loss games indicating that they have sacrificed using their first mover’s advantage for certain games.

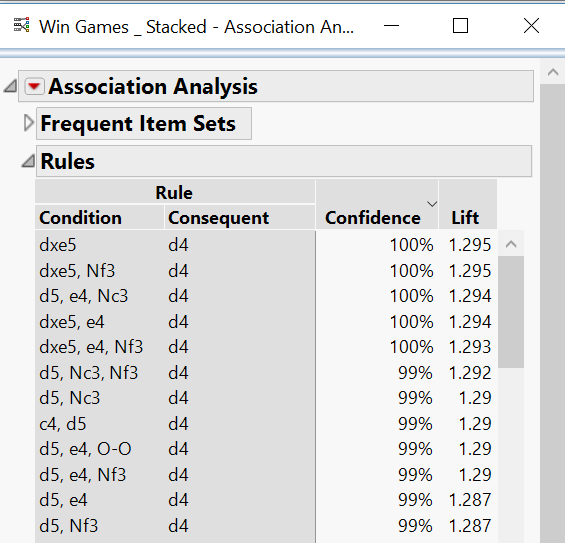


**Figure 4. Association Analysis Frequent Item Sets for Winning Games**

For the loss games, the most frequent opening moves are observed as e4, d4,e3 and Nf3, where a slight deviation in the recommended process is observed where e3 move (**Van't Kruijs Opening**) is used, which does not serve any purpose but instead allows more space for black player to advance into. The queen's bishop's development is slightly hindered by the pawn on e3, and White usually wants to take more than a modest stake of the center.

The suggested best alternative for this step will be e4 as observed in win and loss games but not in loss games.

A dig into the actual chess scenario says that e3 opening move is rarely played and it is considered an irregular opening



**Figure 5. Association Analysis Association rules for Winning Games**

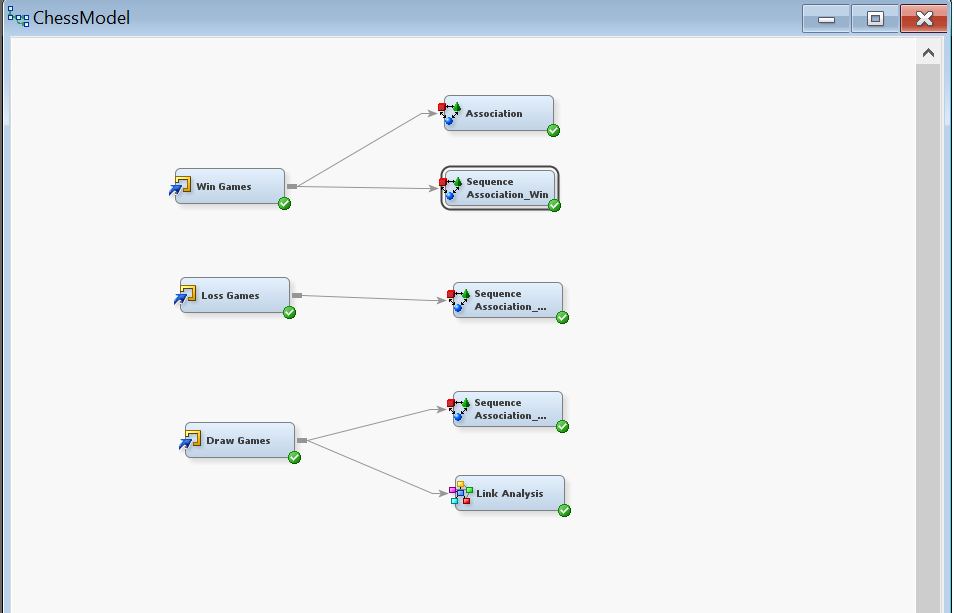
**Lift Factor Observation:**

Bad chess opening choices were not substantiated by recovering measures in the consequent moves in the loss games, leading to further loss of grip in the game and taking him/her even closer to losing.

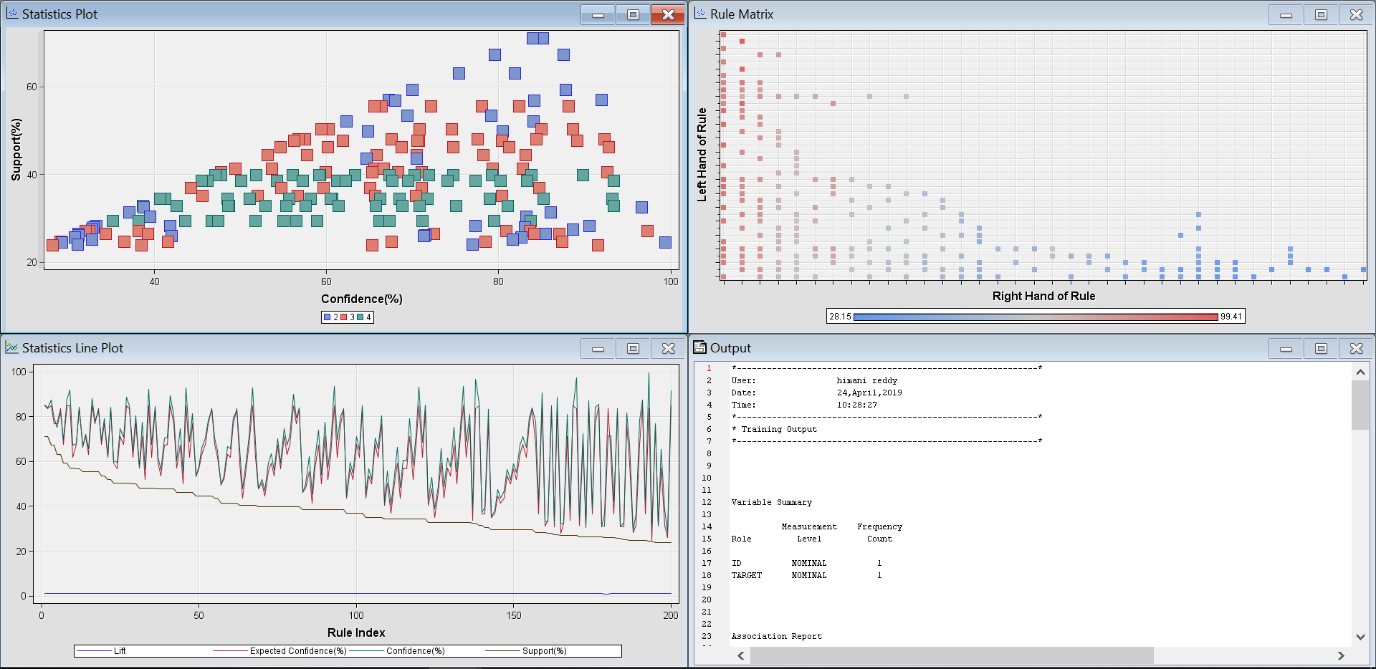
For e.g., In loss games, e3 move is less likely to be followed by c4, d4 moves which might put the player back in the game.

## 6.2 SEQUENCE DISCOVERY

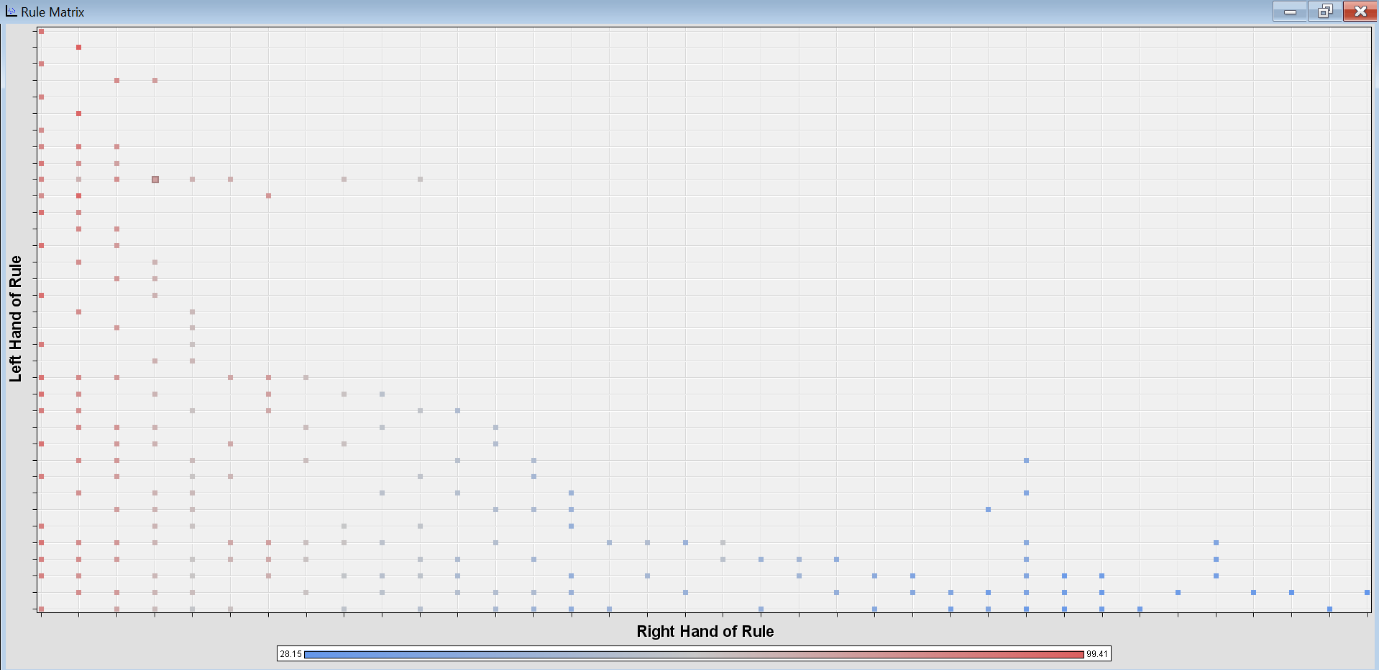
Noting the fact that association analysis alone cannot give us the complete picture of the game as it talks more about the moves occurring in pairs and less about which comes first. Hence, we leaned towards using SAS enterprise miner to accomplish the sequence analysis using Association Analysis node where the sort criteria has been considered as “Support” which specifies the type of support generated for sequence analysis.



**Figure 6. Association and Sequence Analysis Nodes in SAS EM**

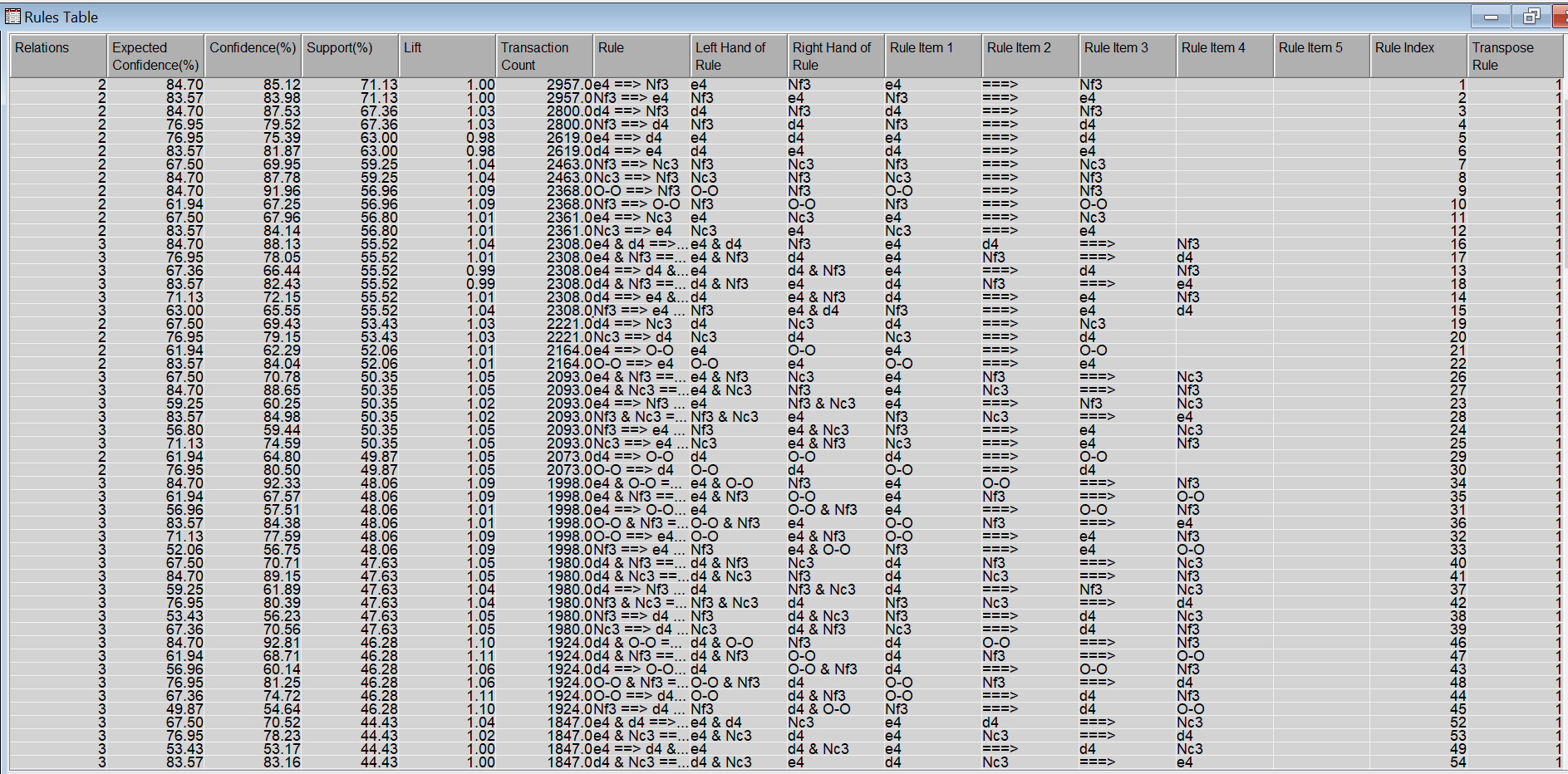


**Figure 7. Sequence discovery results in SAS EM**



**Figure 8. Rule Matrix**

The red squares in Fig 8 shows the move pairs with high confidence. Rules table gives a concise and tabulated form of the most probable sequences enabling you to analyze the final winning sequence and similarly for losing and draw games’ sequences.



**Figure 9. Rules Table for Winning Games**

### 6.2.1 Results

These rules have been analyzed individually to arrive at different sequences. Interesting facts have been observed when the pattern of winning, losing and draw games was studied.

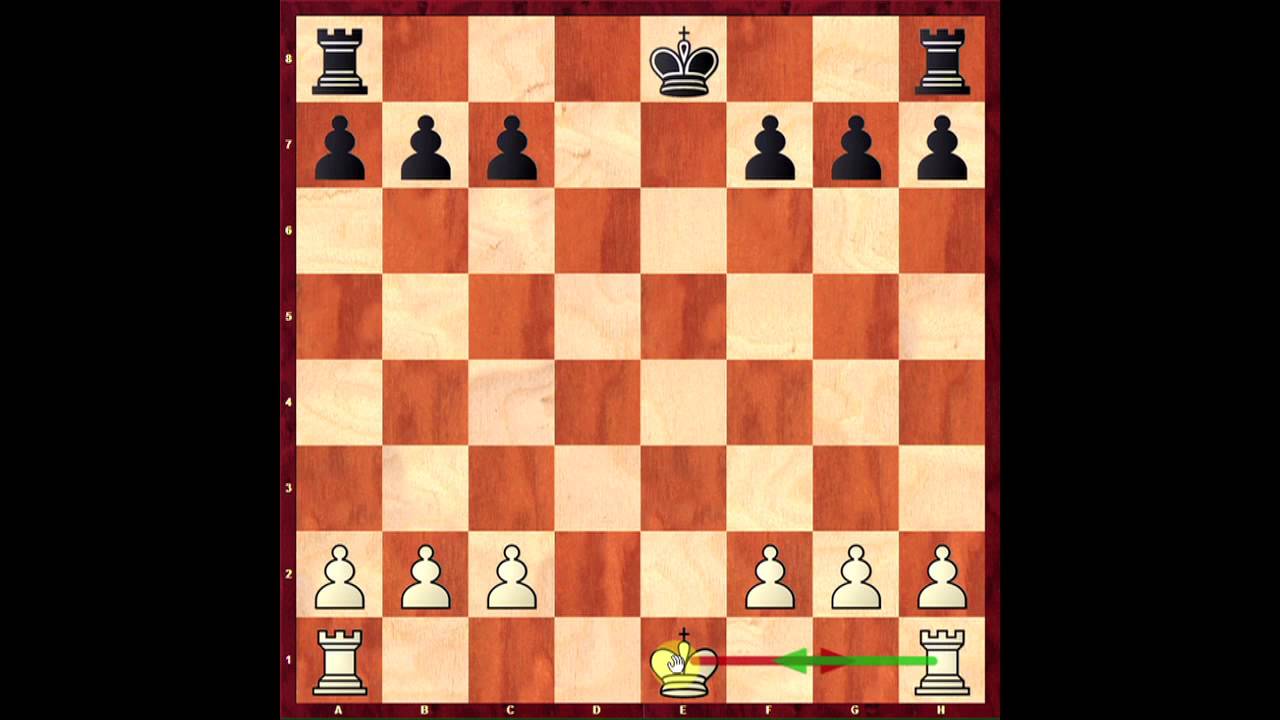
In the loss games, the number of moves taken to arrive at the castling move is more putting them in risk of checkmate.

* The sequence of moves is analyzed ***as e4, d4, Nf3, h3, Nc3, Bc4, O-O*** (Or) ***e4, d4, Nf3, h3, Bc4, Nc3, O-O*** where additional moves such as h3 and Nc3 were made which are not serving any purpose to the goal of safeguarding the King by castling.
* The sequence of win games is observed as ***e4, d4, Nf3, Bc4, O-O***, … which is then followed by additional moves such as Nc3 and c4 etc.

On the contrary the draw games mostly start with **e4, d4, Nf3, Nc3** then leading to various other moves such as c4, h3. There has been no citing of the move **Bc4** which is crucial to the castling move in the first 200 major sequence rules of the all the draw games.

This indicates that the player has taken a long route to castling or castling was not the intention but more utilization of the pieces in the game, leading to more moves in the game and less pieces in the end to checkmate the opposite player resulting in a draw.

**Note: Castling Move**



**Figure 10. King- side Castling (O-O)**

Castling is a special type of chess move where you simultaneously move your king and one of your rooks and only move in chess which involves moving two pieces at the same time. The king moves two squares towards a rook, and that rook moves to the square on the other side of the king. Castling is an important goal in the [opening](https://en.wikipedia.org/wiki/Chess_opening), because it serves two valuable purposes: it moves the king into a safer position away from the center of the board, and it moves the rook to a more active position in the center of the board.

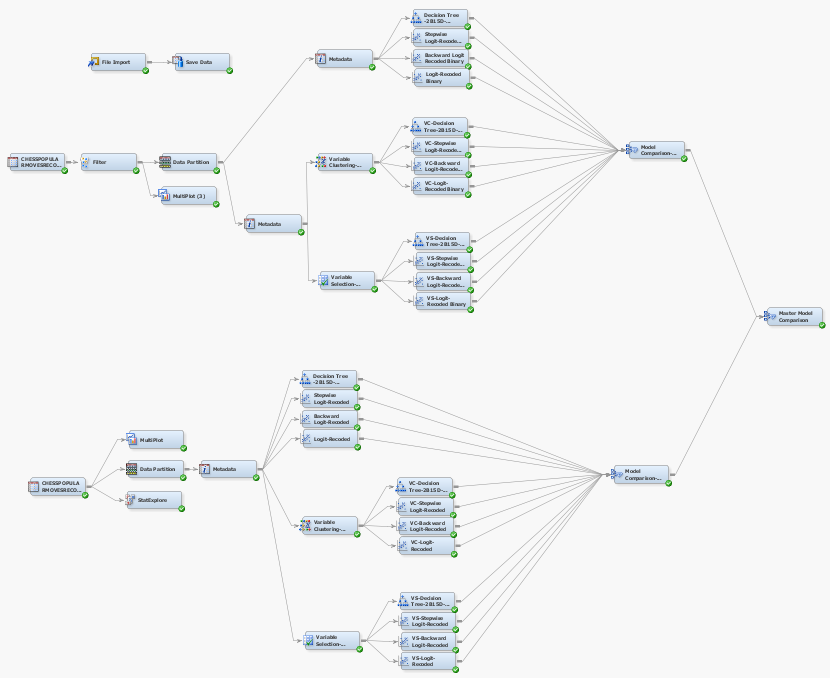
# 6.3 BINARY AND MULTINOMIAL LOGISTIC REGRESSION

## 6.3.1 PURPOSE

Logistic Regression is done to understand the effect of moves, as well as the effect of frequency of each move on the game’s outcome.

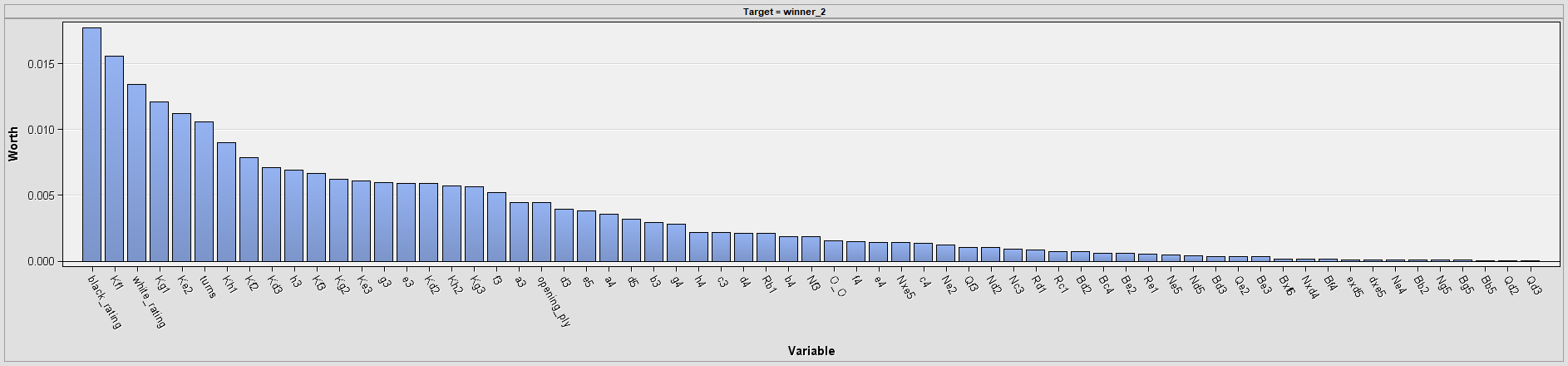
## 6.3.2 PROCEDURE AND RESULTS

Data is imported using file import and unnecessary variables were rejected. 61 moves identified as being popular, along with *turns*, *opening\_ply*, *white rating* and *black rating*, and the target variable which is *winner\_2* were included.



**Figure 11. Model Generation and Comparison**

* Variables were analysed for quasi and complete separation using ‘Multiplot’.
* The ‘filter’ node was used to manually filter out the ‘draw’ cases for use in binary regression.
* Data Partitioning was done in the ratio of 2:1 for training: validation data.
* Variable rejection was done in multiple stages using metadata node. The variables finally used in the models were all high in variable worth.

****

**Figure 12. Variable Worth**

* All the models generated are passed through intermediate model comparison nodes and finally onto Master Model Comparison node, which identified Stepwise Binary Model to be marginally better in terms of ROC index as well as Misclassification Error.

### 6.3.2.1 BINARY LOGISTIC REGRESSION

Binary logistic regression was carried out to compare win against loss for all the predictors, with win being the ‘comparison group’ and loss being the ‘referent group’.

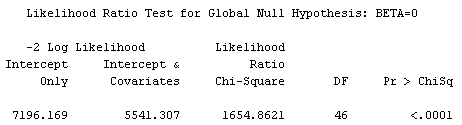
The odds ratios for multinomial regression are interpreted as follows (Multinomial Logistic Regression| SPSS Annotated Output):

An odds ratio > 1 indicates that the comparison outcome is more likely. An odds ratio < 1 indicates that the risk of the outcome falling in the comparison group relative to the risk of the outcome falling in the referent group decreases as the variable increases. In other words, the referent outcome is more likely.



**Figure 13. Binary Logistic Regression – Stepwise - Selected Model**

**Whole Model Test**

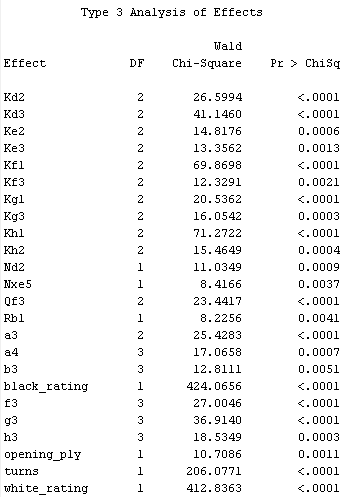


**Figure 15. Binary Logistic Regression - Likelihood Ratio Test**

The selected model passes the whole model test with a large chi-square value and a significant p-value of <0.0001, proving that at least one of the regression coefficients for the explanatory variables is non-zero.

**Type 3 Analysis of effects**

This test is carried out to see whether the parameters included in the model contribute to improving the model fit. Despite building a stepwise regression model, there were predictors with very high p-values, which made them insignificant to the model. Such insignificant parameters were systematically removed using metadata node.



**Figure 16. Significant model variables seen in Type 3 analysis of effects**

**Analysis of maximum likelihood estimates, and Odds ratio estimates**  
The maximum likelihood estimates and odds ratios from the SAS Report were combined in MS Excel and were filtered to include only parameters and their levels that had p-value<0.05 for estimates.

**Moves and Factors which increase propensity of winning**



**Figure 17. Predictors and their levels that increase odds of winning**

The following are the moves taken from the above screenshot, which increase odds of winning (vs losing):

1. Moves that should not be made often to improve odds of winning are listed in descending order of their ability to influence below:
   1. *Kd3, Kh1, Kg3, Kg1, Kd2, h3, Kf1, Kf3, a3, a4, Ke2, Kh2, Ke3, Nd2, Rb1, b3, g3*
2. Moves to be made once instead of more times to increase odds of winning, listed in descending order of their ability to influence a win:
   1. *h3, f3, a4*
3. Other factors influencing win:
   1. *Higher opening play moves, higher number of turns and higher white rating*

**Moves and Factors which decrease propensity of winning**

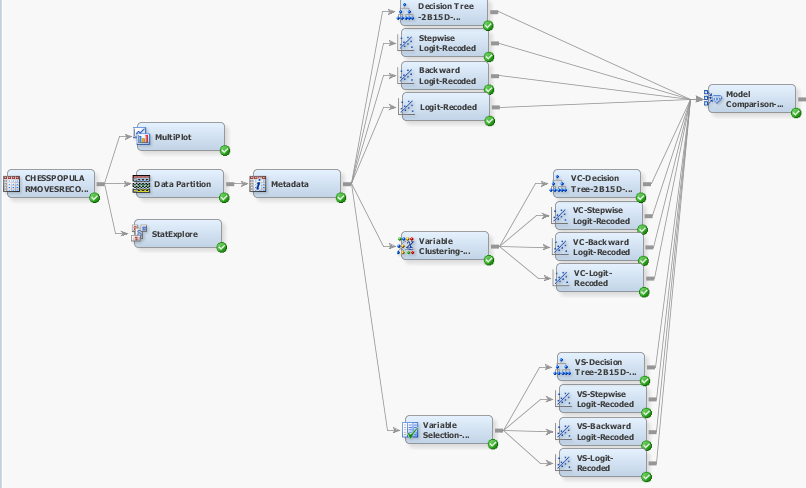
****

**Figure 18. Predictors and their levels that decrease odds of winning**

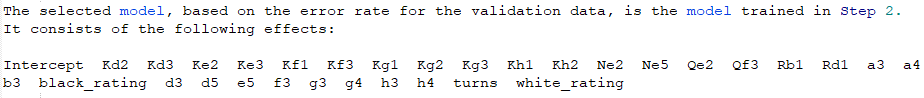
The following are the moves taken from the above screenshot, which decrease odds of winning (vs losing), or, increase odds of losing:

1. Moves which, when not made, increase odds of loss:
   1. Qf3 (making 2 and above moves reduces odds of losing), Nxe5 (making 1 and above moves reduces odds of losing)
2. Moves which, when made once instead of more times, increase odds of loss:
   1. g3 (making 3 and above moves reduces odds of losing), Kf1(making 2 and above moves reduces odds of losing)
3. Other factors influencing loss:
   1. Higher black rating

### 6.3.5 MULTINOMIAL LOGISTIC REGRESSION

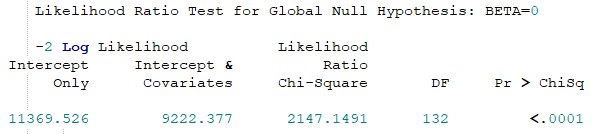


Multinomial regression, in the context of this analysis, makes k-1 models in general, where k=3 here, as that is the number of levels (Win/Loss/Draw) of response variable. Therefore, it makes 2 models, one with each of the 2 comparison outcomes (Win or Loss), vs one referent outcome (Draw).



**Figure 19. Multinomial Logistic Regression – Backward - Selected Model**

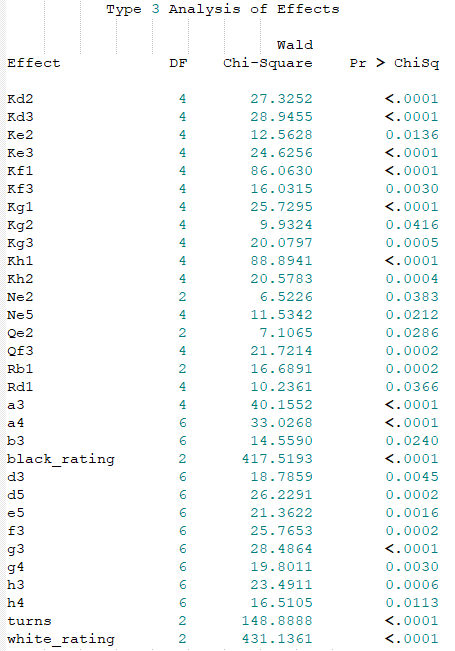
**Whole Model Test**



**Figure 20. Multinomial Logistic Regression - Likelihood Ratio Test**

**Type 3 analysis of effects**

Despite building a backward regression model, there were predictors with very high p-values, which made them insignificant to the model. Such insignificant parameters were systematically removed using metadata node. The remaining parameters are all significant with p-value <0.05 and contribute towards improving the model fit.



**Figure 21. Multinomial Logistic Regression - Type 3 Analysis of Effects**

**Analysis of maximum likelihood estimates and Odds ratio estimates**

#### Sub-model: Win vs Draw

The maximum likelihood estimates and odds ratios from the SAS Report were combined in MS Excel and were filtered to include only parameters and their levels that had p-value<0.05 for estimates.

1. **Insights - Factors which increase propensity of winning (vs draw)**



**Figure 22. Multinomial Logistic Regression - factors which increase odds of winning vs draw**

|  |  |  |
| --- | --- | --- |
| **Winning (vs Draw)** | | |
| Moves that should not be made to improve odds of winning | Moves to be made once or less to increase odds of winning | Other factors |
| Kg3 | h3 | Higher white rating |
| Kg1 | a4 |  |
| Kh1 | f3 |  |
| Ke3 | g4 |  |
| Kh2 | d5 |  |
| Kf3 |  |  |
| Kd3 |  |  |
| Kf1 |  |  |
| Kg2 |  |  |

**b. Insights - Factors which decrease propensity of winning (vs draw)**



**Figure 23. Multinomial Logistic Regression - factors which decrease odds of winning vs draw**

#### Sub-model: Loss vs Draw

1. **Insights - Factors which increase propensity of loss (vs draw)**



**Figure 24. Multinomial Logistic Regression - factors which increase odds of losing vs draw**

The following moves increase odds of losing (vs Draw):

1. Moves to be avoided to reduce odds of losing, listed in descending order of their ability to influence a loss:
   1. Kg1, Kh2, d5, g4, Kh1, g3, f3
2. Other factors influencing a loss (marginally):
   1. Higher black rating
   2. **Insights - Factors which decrease propensity of loss (vs draw), or increase the probability of match ending in a draw**

****

**Figure 25. Multinomial Logistic Regression - factors which decrease odds of losing vs draw**

1. Moves which, when never made, reduce odds of loss (increase odds of draw) than when they’re made:
   1. a3, Kh1, a4, Kf1, Rb1, Kd2, h3, Qe2, d5
2. Moves which should never be made 3 or more times, in order to reduce propensity of loss:
   1. a4, g3, h4 (making 3 and above moves of h4 reduces probability of draw, while making this move twice comparatively increases probability of draw)
3. Other factors reducing odds of loss (and increases propensity of draw):
   1. Higher number of turns, higher white rating.

**OVERALL INSIGHTS FROM BINARY AND MULTINOMIAL LOGISTIC REGRESSION**

****

**Figure 26. Overall Insights**

1. Moves involving king's movement such as Kd3, Kh1, Kg3, Kg1, Kf2, Kf3, Ke2, Kh2, Ke3..., when moved once or less, are observed to increase the odds of winning. This, in lay man terms, translates to the idea that the King's movement ought to be kept to a minimum, unless there is a dire need or when he is under checkmate
2. Also, the opening moves such as a4, g4, h3, f3 are found to not contribute much to the win of a game as they don't seem to be utilising the white's advantage, which is to be able to attack the centre of the board to have a lead in the game

# 7. CONCLUSION

In conclusion, being a game of patterns, dataset posed an immense and interesting challenge to analyze and decode. Summing up the analysis, when a player is playing on the white side, he/she is advised to make use of first movers’ advantage and occupy the center position on the chess board. As, deduced from the explanatory analysis using certain opening moves such as Sicilian Defense which are well acclaimed and most preferred will increase the odds of winning. Aiming to castle the king as soon as possible shall also further increase the probability of winning as mentioned by the ‘winning sequence’ from the sequence discovery analysis. The winning sequence is found to be e4, d4, Nf3, Bc4 and O-O. Another important take away from the analysis is to keep your King’s movement to minimum as advised by the insights of Logistic regression.

# FUTURE WORK

In this analysis, white player’s moves and strategy are studied and analyzed. Future work could include the analysis of black moves and compare the two analysis to further understand how a countermove affects the progress of the game. What best countermoves shall lead a player to success and how to predict the opponent’s strategy and predict his next move in advance analyzing the moves he/she had already done.

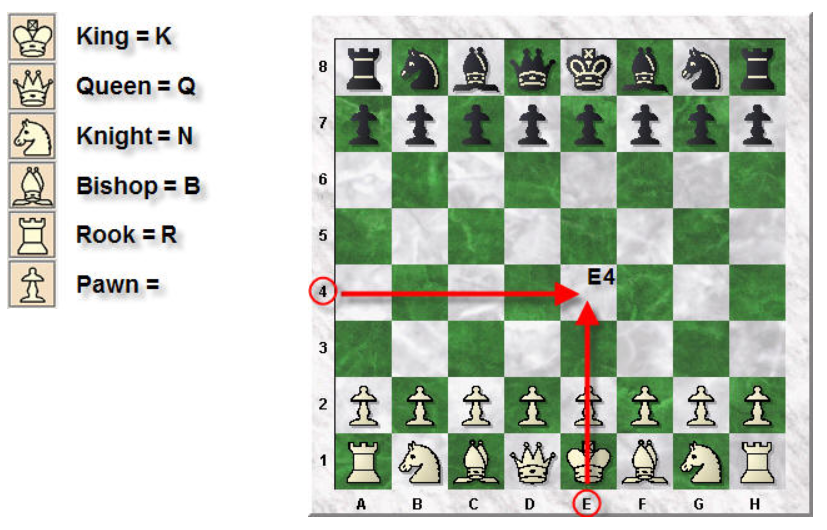
# 8. REFERENCES

Ed. (2019). *There’s Nothing Odd about the Odds Ratio: Interpreting Binary Logistic Regression*. Retrieved April 2019, from Statistics Solutions: https://www.statisticssolutions.com/theres-nothing-odd-about-the-odds-ratio-interpreting-binary-logistic-regression/

*Multinomial Logistic Regression| SPSS Annotated Output*. (n.d.). Retrieved April 2019, from UCLA Institute for Digital Research & Education: https://stats.idre.ucla.edu/spss/output/multinomial-logistic-regression/

Raphael. (2017, March 1). *How to Read and Write Algebraic Chess Notation*. Retrieved April 2019, from ChessHouse: http://blog.chesshouse.com/how-to-read-and-write-algebraic-chess-notation/

# 9.APPENDIX 1. Algebraic Chess Notation (Raphael, 2017)



# 10. Acknowledgments

First and Foremost, we would like to show our immense gratitude to our instructor and mentor, Associate Professor Kam Tin Seong, for sharing his pearls of wisdom which enabled us to bring our project to completion.

We would also like to thank our friends and fellow students, whose insightful discussions sparked ideas that made the project what it is.

# 11. Recommended Reading

# Contact Information

Your comments and questions are valued and encouraged. Contact the author at:

Name: Alakananda Giridhar

E-mail: giridhara.2018@mitb.smu.edu.sg

Name: Himani Reddy

E-mail: himanimukka.2018@mitb.smu.edu.sg

Name: Rahi Sheth

E-mail: rahi.sheth.2018@mitb.smu.edu.sg

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.