**The Road on Becoming the Next E-Sports Professional**

By Edward Lee

The world of E-Sports is growing at a very rapid pace and many young prospects are looking towards gaming as a potential career choice. However, how does one person decide if their talent or skill level is enough to compete with the best in the world? How much time should be invested or achieved to be better than an average player? What is the threshold to decide whether one person is suitable for this career choice? What is the difference between playing the game myself and the professional players I watch on twitch.tv or other online live streaming websites?

The dataset I will be using is the SkillCraft1 Master Table Dataset Data Set from the University of California, Irvine Machine Learning Repository dated on September 20th, 2013. The dataset is based on StarCraft II replay files received from players in the lowest league to the players competing at a professional level. The replay files are from 1v1 matches running StarCraft 2 version 1.3.6.19269:291.

The analysis is not targeted to any company or journalist because I am trying to present that reaching a level to compete against the best players in the world, is not easy and actually requires dedication. My target audience is towards all younger generation considering professional gaming as a future career. Based on my analysis, I hope to enlighten them on how much minimum time should be invested to be able to compete against the best in the world and, if they are talented, what minimum skills must be achieved. Without the knowledge, the younger generation will never know the difference between their own talents versus the world professionals already in the business.

The dataset includes many interesting variables, but the most interesting variable is the APM or Actions per Minute. APM is simply 1 action of every click of the mouse onto an object in the game that produces an immediate result or the hitting a key on the keyboard to switch frames, objects (units, structures or spells), to build or upgrade a structure or to produce a unit. In theory, when a player’s APM is much higher than their opponents, the player with the higher APM should win because they would be able to gather more resources, produce more structures and units and command their army on their will. Through the dataset, the highest achieved sustained APM through an entire game is 389.80 with a median of 108 and mean of 117. The difference between an average player and, most likely a professional player, is a huge difference. However, just slightly glancing through the LeagueIndex of 8 (professional players), there are also APMs as low as roughly 146 and 179. Does APM really stand between a potential youngling looking into becoming the next E-Sports professional or is it just a number to scare you away?

However, after learning and educating myself on the other variables, I have noticed that there is an even more important variable that may trump APM, although not by a lot. The PAC (Perception-Action Cycle) variables, which include NumberOfPACS, GapBetweenPACS, ActionLatency and ActionsInPAC, may be the most important variables that can differ an average player to a professional player. PAC, in StarCraft II terms, is the shifting of a screen, then followed by at least one action before shifting to a new location on the screen. Many may ask, why are these PAC variables more important than APM? Well, APM has flaws. Although it is still a very good indicator for how many actions a player can manage per minute, it is too generalized. Players who are very good at macro-management, which is the constructing and producing of structures and units, may not be good at micro-management, which is moving your units, attacking with your units and using of unit abilities, and vice versa. Every key pressed or mouse click adds to 1 APM, but that doesn’t show what has actually been done. Players may have bound numbers 1 to their main structure and 2 to their barracks, press the keys over and over again, and still produce a very high APM, which is not accurate. However, PAC calculates the actions a player has done after shifting screen, which can more accurately show that a player has continuously changed their perspective on the game to, not only build their army and structure up, but to adjust to the game at real time. Players who are able to constantly, within low durations (GapBetweenPACS), perform high amounts of screen shifting (NumberOfPACS), high amounts of action per each shift (ActionsInPAC) and the milliseconds before their perform an action (ActionLantency), can arguably be concluded to be a better player and have a better understanding of the game.

However, there are some limitations to the dataset provided. There is no Win/Loss column, Win Rate column or Win rate against higher or lower ranked players. Also, it would be interesting to have a mean length per game for each rank to see if there are any disparities between low ranking games versus high ranking games. Also, we do not know which country of origin the professional players are from versus the players who submitted their data. The overall quality of players in South Korea are much stronger than the overall quality of players in North America and Europe. It would be good to know where the data came from.

The dataset itself is fairly clean, with a couple of minor exceptions and easy to deal with issues. LeagueIndex = 8 players (professional players) have empty cells in Age, HoursPerWeek and TotalHours. The age problem was fixed by taking the mean of each column and inputting the mean value into each cell. As for HoursPerWeek and TotalHours, I took the game release date subtracted from the date the dataset was released, estimated that professional players play roughly 8 hours per day (normal working hours for any white-collar employee) and inputted those values accordingly. There are only 55 players at LeagueIndex = 8 out of 3395 instances, so it was not going to skew the data by a wide margin. Also, the data recorded in the dataset was recorded while on fast, which is 1.5x speed the normal speed. Therefore, the dataset used timestamps instead of seconds. The dataset provider listed that 1 real time second equates to 88.5 timestamps. I switched the timestamps into seconds for easier calculation for seconds, minutes and hours.

Preliminary exploration of the dataset shows that the mean and median, as well as the range, increases as players progressed. However, the difference in mean and median from GrandMaster, the highest rank achievable online, and professional players has a much larger gap than from Master to GrandMaster. PAC is much more interesting because the gap is not as big as I thought. Although there is still a difference between GrandMaster and professional players, we can see that GrandMaster players do have the ability to compete with professional players. However, this requires more detailed data gathering, if there data is even out there. Also, I would like to point out that I would imagine that the majority of professional players will mostly be practicing online, rather than against other professional players on a daily basis, which may skew the actual data of GrandMasters. However, this is just an assumption and without more detailed information about where the data came from, the conclusion is null and void.

I have also noticed a trend that plotting the median age (age 21) versus the median hours played (12 hours per week) and colored by the rank of the player, players who played much more were more likely to be at a higher rank. However, what caught my attention was, LeagueIndex = 6 (Master) and LeagueIndex = 7 (GrandMaster), were the only 2 ranks much higher than the rest. From starting out as a newbie (LeagueIndex = 1/Bronze) to becoming a good player (LeagueIndex = 5/Diamond) does not require extra time, rather it may be the difference of talent. However, becoming a good player to a top player online, may require the player to prioritize more hours to the game.

Based on the current findings and deeper understanding of the variables, I will be exploring the correlation between variables much more. There may be more interesting relationships between variables that I may not have seen, which I am very interested to delve into. However, my main calculation will still be the Epsilon Regression method through SVM (Support Vector Machine) in R and using the RMSE (Root Mean Squared Error) and Cross Validation Error with 5 folds.