Comp 4441 Final Project

Shane McIntyre

Grouping Player’s Skill Level by Kmeans

One of the most popular video game franchises of all time is Call of Duty. Through controversy, age play, and other factors, the franchise has proven to be one of the most influential gaming franchises over the last 2 decades. As a gamer myself I wanted to test how statistics can be an important factor in a game like call of duty, or if the phrase “RNG” for gaming is true.

RNG in the data world means random number generator. However, in gaming the phrase is used to refer to the randomness of other players activity in a game impacting how successful you are.

Growing up people who played lots of video games were seen as nerdy, but in today’s world that is not the case. Many professional athletes and celebrities love to game, and streaming platforms such as Twitch have grown very popular. Every kid at some point has wished to be a streamer, and in today’s world streaming video games has become popular. The question is, how common is it to consistently perform at an elite level where people would want to watch your gameplay? I found a data set on Kaggle that collected leaderboard data for Call of Duty: Modern Warfare (2019). Arguably the most important statistic is Kill Death Ratio (kdRatio or KD), i.e. how many times you kill a player versus how many times you get killed. In a perfect world, having a KD of 1.0 would make sense, having 1 kill per every death. However, based on the data set this is not true.

**Data Processing:**

To begin, I took the data set and filtered it down to players who had played more than 100 games. There was a fair number of players who played a handful of games and then quit, so I wanted to remove those players. This cut the data set down from 1558 players to 403. While 403 observations are not a lot, when wanting to look at how likely it is to be consistently great at the game, I only wanted to compare players who played consistently.

The next step was to see how the recorded game statistics interacted with each other.

Chart, scatter chart

Description automatically generated

From the correlation plot some basic understanding can be seen. A player with more kills will win more often and has a higher in game level and time played. A player with more deaths had a lower kdRatio. These basic comparisons are easy to spot out, but when looking at Kill Death Ratio (kdRatio in the plot), the correlations are much more difficult to see.

Since I was interested in the KD stat, I isolated that statistic to get a better understanding of the data. It turns out the average KD of the data set was less than one, which can be seen below:

Chart, line chart

Description automatically generated

The average KD was 0.9768, with an upper and lower quantile of 0.8818 and 1.0468 respectively. From this, I grouped KD into four main groups:

* Bad is a KD <= 0.8818
* Average is a KD between 0.8818 and 1.0468
* Good is a KD between 1.0468 and 1.5
* Elite is a KD better than 1.5

Since the distribution of kdRatio appeared to be normally distributed, I wanted to test if that was the case. I plotted out a qq plot to check the normality visually:

Chart, line chart

Description automatically generated

While the bulk of the data appears to be normally distributed, the ends tailed off more aggressively than a true normal distribution would. My next step was to do a statistical test. Since I was looking at just kdRatio, I did a Shapiro-Wilk’s test of normality. The test held a p-value < 2.2e-16, telling me the data was not a true normal distribution. Originally, this led me to scale the data set, but a scaled data set had a (between\_SS / total\_SS = 57.9 %) while unscaled was (between\_SS / total\_SS = 90.4 %) when going through the KMeans algorithm.

**K-Means Clustering:**

The next step was to set up the Kmeans algorithm. I decided to use 4 clusters (k = 4) since I grouped the kdRatio into 4 categories (Bad, Average, Good, and Elite). Given the complete data set, the KMeans algorithm didn’t group the data into the 4 skill level groups too accurately. My take aways from this will be discussed below. While my 4 predetermined groups did not match to well to the groups the algorithm found, when grouped against all 17 in game statistic variables, the algorithm was able to group into its own 4 distinct groups well.

Chart, scatter chart

Description automatically generated Chart, diagram, bubble chart

Description automatically generated

As you can see there is some blending between the groups, but overall they have their own set of boundaries.

Since my confusion matrix didn’t match well to my predetermined skill groups, I decided to test one more time. The difference this time was I took the 2 variables with the highest correlation to kdRatio, which was Headshots (numbers of head shots a player landed on others) and Level (the in game level which is raised by good gameplay or a lot of time played). The resulting plots are as follows:

Chart, scatter chart

Description automatically generated Chart, diagram, bubble chart

Description automatically generated

**Take Aways:**

While the Kmeans algorithms didn’t group well into my predetermined skill groups, it was still able to locate a trend and make four relatively distinct groupings. The first of the 2 plots appears to have some form of linear relation between headshots and level and the groupings, which leads me to believe there is a way to use in game statistics to classify players based off of kdRatio, but an in depth linear regression model might be a better tool for this. Additionally, the cluster plots do show some blend between the groupings, which leads me to believe that video game “RNG” heavily impacts how a player performs. Outside of the elite players who always perform at the highest level, in an online first person shooter game there are times where a bad player can perform as well as a good player depedning on how the game plays out. In other words, it is not too uncommon for a bad player to have a great game, or a great player to have a bad game. My next step in this analysis will be to transition from KMeans to a linear regression model where kdRatio is dependent on the other 16 variables, and try to formulate an equation to estimate a players kdRatio or skill level.