Coby Lin

12 May 2025

Data Bootcamp 001

Koehler

Worlds 2024 Champion Data Analysis and Regression

For this project, I conducted an exploratory data analysis and created regression models using a dataset on the 2024 League of Legends World Championship from Kaggle. The dataset I used contains information on champion names, the number of times they were picked and banned, overall presence percentage, wins, losses, win rate, kill-death-assist ratio (KDA), average build time, total game time, creep score per minute (CSM), damage per minute (DPM), gold per minute (GPM), and three other differential statistics.

For reference, each character within League of Legends is referred to as a champion. Each feature in the dataset demonstrates a portion of a champion's overall power within the overall current META (Most Effective Tactics Available). This refers to the most powerful team composition in the game for professional play. For example, the number of picks refers to how many times the champion was picked by a team throughout the tournament. The number of bans shows how many times a character was banned from a game by either side of a match. The presence percentage shows the collective pick and ban percentage compared to the total number of games in the tournament. The wins and losses columns refer to the number of games that a champion won or lost. Winrate refers to the percentage of games won compared to total games played by the champion. The kill-death-assist ratio refers to how many kills and assists on enemy champions were scored by a particular champion divided by the number of deaths that the champion had. Average build time refers to how long a champion takes to build one of their core

items, enabling them to become stronger throughout the course of the game. Total game time refers to the number of minutes that a character was in play for over the course of the entire tournament. Creep score per minute (CSM) refers to the number of minions/monsters collected within a game. Damage per minute (DPM) refers to the mean amount of damage a champion outputs each minute of a game. Finally, gold per minute (GPM) is the mean amount of gold that a champion earns per minute of game time.

In this project, I aimed to examine what factors led to the most champion presence in the championship. To achieve this goal, I created scatterplots and a line plot (for average build time) for each factor and their correlation to presence. When creating the plots, I made sure to filter all the data based on champions that were picked at least 10 times over the course of the tournament. This made sure that the graphs weren't heavily skewed due to using data with low sample sizes. Based on the exploratory data analysis alone, I concluded that Average Build Time and Wins had the greatest impact on the presence. Both of the scatterplots for these factors showed a strong correlation to presence, with wins having a direct correlation while average build time had an inverse correlation. Based on the strategy behind the game, these correlations make sense. The more wins a champion is able to acquire, the more attractive an option they are to be played in professional games. Meanwhile, the less time it takes for a champion to build one of their core items, the sooner they can become powerful and contribute to a win. I also examined the effects of picks and bans on presence. While both of these factors had a strong direct correlation with presence, it would be meaningless to include them since they directly contribute to the presence percentage. The other factors in the dataset did not result in any significant correlation with presence.

Since the dataset contains exclusively numbers that could not be classified, I built three different regression models to test my findings. In particular, I utilized linear regression, KNN regression, and RandomForest regression. For the linear regression model, I first use the average build time and wins columns to predict presence. Similar to the exploratory data analysis, I made sure to drop the rows where champions were played less than 10 times to prevent skew. I also made sure to inverse the values of the average build time column to reflect the inverse correlation between the two variables. I then created a pipeline with a StandardScaler to scale the data and then perform the linear regression. From that model, I got a 45.6% score and a 0.023 mean squared error (MSE). For comparison, I took the mean of the presence column to provide a baseline, which was about 0.534. Since the model's MSE is lower than the baseline, it suggests that the model has a good amount of predictive power. I also experimented with several other variables, such as losses, DPM, GPM, and CSM, to see if they would increase the score or lower the MSE, which ended up not happening. However, when I ended up adding the KDA column, the score jumped to 48.6% and the MSE went to 0.022. This suggests that KDA might be a good predictor of presence. Next, I used a KNN regression model using the same two factors initially. This ended up giving me a score of 47.2% and an MSE of 0.022. After adding in the KDA column into my model, the score increased to 70.8% and the MSE went to 0.012. This significant jump further suggests KDA as a good prediction factor of presence. Once again, I also tested a few other variables, which lowered the accuracy of the model. Finally, I used a RandomForest regression model for my findings. I started out with the same two columns as the linear regression model. Using those two factors alone, I ended up getting an 18% score and a 0.035 MSE. However, after using KDA once again, the score jumped to 22.2% and the MSE went to

0.033. When testing this model, I also experimented with other variables, which did not result in any improvements.

Although KDA in the EDA did not reveal any significant correlation to presence, the three models all improved when adding in the column. This strongly suggests that KDA is a strong indicator of presence. In the context of the game, this finding makes sense. If a champion is able to get a high amount of kills and assists with few deaths, that is an indication that the champion is strong in the current META. Therefore, we can conclude that the three most important factors to determining champion presence in a professional League of Legends game is the number of wins, the average build time, and the KDA. While the initial exploratory data analysis suggested a strong correlation between Wins and Average Build Time with Presence, the regression models revealed the significant predictive power of KDA, even though its direct correlation in the EDA was less pronounced. This highlights the ability of regression models to discover more complex relationships between variables. Among the models tested, the KNN regression model incorporating KDA performed the best, achieving a significantly higher score compared to linear regression and the RandomForest models. The improved performance of the KNN model suggests that non-linear relationships may exist between the champion statistics and their presence in a tournament.

For further analysis, it might be beneficial to test other types of regression models to explore the effectiveness of these three factors. Both the linear regression and RandomForest models showed a sub-50% model score, indicating that there is significant work that can be done to improve them. For the RandomForest model in particular, it might be beneficial to add in a GridSearchCV to automate the process of finding optimal estimators, max depth, etc. to improve accuracy. For a game as complex as League of Legends, publicly available data on patch

changes, professional player tier lists, or meta analyses could provide valuable context and potentially be incorporated as additional features. There are also game features that could be incorporated such as champion class, item costs, and synergy with team compositions. Exploring how champion primary roles (e.g., Tank, Mage, Marksmen) correlate with presence could be a gateway into more complex synergy analysis. However, incorporating these more nuanced variables would require significantly more advanced modeling skills. It is also important to consider the different kits and play styles that each champion has, which is difficult to measure out in a predictive model.