Big Data and Predictive Analysis Capstone Project: An Empirical Study of League of Legends Solo Queue Games

**Introduction**

League of Legends is a multiplayer online battle arena(MOBA) game played by millions of players of varying demographics. Attempting to quantify the factors that have contributed to the game’s growth is a convoluted task. However, two characteristics are vital not only for sustained participation but are also relevant features that allow League of Legends to be studied empirically for the purposes of this project.

1. League of Legends is a game of inherent and acquired skill
2. Theoretical Novelty is achieved within every game

These topics will be explored further in the literature review section as well as the final report. However, it is important to distinguish these features from other characteristics of the game such as emotional and aesthetic appeal.

While there are several game modes offered to League of Legends players, the focus of the study will be observing Solo Queue Ranked game stats played on the map Summoner’s Rift. This game mode offers users the option of queuing for games with 9 other human participants divided into two teams through a process called Champion Select. The primary goal of this game mode is to destroy the opponent’s Nexus through the successful application of a variety of strategies and techniques. Some of these strategies or techniques have widely accepted game theory optimal methods but complex metagame strategies also exist which are either not defined or difficult to quantify.

The primary purpose of this project is to quantify the variety of win conditions from a series of available in game meta data available through the Riot Games API. Statistical techniques will be applied to identify strength of correlation between individual player performance and win/loss result. League of Legends provides an interesting opportunity for a supervised learning experiment due to the availability of extractable continuous variables and a dichotomous supervisory signal. Existing methods of performance evaluation such as KDA(Kill Death Assist ratio) and kill participation will be analyzed as it relates to the dichotomous ground truth of winning or losing. Feature selection, correlation analysis, and prediction modelling will be some of the tools used to accomplish these tasks.

League of Legends is generally perceived to have a steep learning curve to play at a high level and a certain level of domain expertise is required to comprehend the purpose of this project. A significant but ancillary task for this project is also to transcribe some of these components of the game to pupils familiar and foreign to League of Legends.

**Literature Review**

Complete lack of formal research on the topic of League of Legends is both a constraint and an advantage for the task of creating a novel project. Websites and services currently exist which allow users to analyze certain aspects of their performance. Some of these services will be evaluated at various points throughout the duration of the project but they do not provide a framework for comparison. However, there is a branch of academic research which I can draw inspiration from and that is the study of complex systems. So what is complex systems theory and why is it relevant to this project?

As described by Yaneer Bar Yam in *Dynamics of Complex Systems*:

*The objectives of the field of complex systems are built on fundamental concepts— emergence, complexity—about which there are common misconceptions…*

*A complex system is a system formed out of many components whose behavior is emergent, that is, the behavior of the system cannot be simply inferred from the behavior of its components. The amount of information necessary to describe the behavior of such a system is a measure of its complexity.*

*We see that emergent properties cannot be studied by physically taking a system apart and looking at the parts (reductionism). They can, however, be studied by looking at each of the parts in the context of the system as a whole. This is the nature of emergence and an indication of how it can be studied and understood.*

*If we think about the system as a whole, rather than the small part of the system, we can identify the system that has a global emergent property as being formed out of interdependent parts. The term “interdependent” is used here instead of the terms “interconnected” or “interwoven” used in the dictionary definition of “complex” quoted in Section 0.1, because neither of the latter terms pertain directly to the influence one part has on another, which is essential to the properties of a dynamic system. “Interdependent” is also distinct from “interacting,” because even strong interactions do not necessarily imply interdependence of behavior.*

*The second concept that is central to complex systems is a quantitative measure of how complex a system is. Loosely speaking, the complexity of a system is the amount of information needed in order to describe it. The complexity depends on the level of detail required in the description.*

It is not the goal of this project to prove that League of Legends can irrefutably be defined as a complex system. It is however, a branch of academia which provides some insight into how a complex system can be analyzed. Can the dynamically shifting metagame strategies in League of Legends be considered an emergent property within the context of complex systems? Can the finite rules that govern behaviour within the game be considered analogous to the rules that govern natural systems? How does theoretical novelty impact the ability to quantify interdependence in a complex system? These are some of the philosophical questions I will continue to wrangle with throughout the duration of the project.

**Approach**

The approach section has been divided into two sections. The first block diagram will attempt to compartmentalize the programming steps and the second block diagram will outline the approach to analysis.

**Step 1: Accessing Riot Games API**

The first step to acquiring access to the Riot Games API is to login using a Riot Games account and registering for a unique API key. This was done prior to the start of the project. There is a limited amount of official and unofficial support documentation. A Python script must be created to access the match stats of a game. The Python script must do several things to compile the dataset.

The RIOT games API allows for a limited number of API calls per API key. This can be a potential issue when attempting to make enough API calls to create a sufficiently robust dataset. The sample size of the dataset will be somewhat gated by this limiter but there are work arounds to this issue such as requesting a API call limit increase or using multiple API keys using multiple accounts.

**Step 2: Acquiring the Match List**

Due to a recent change in Riot Games API policy, a list of matches must be compiled using player seed information by username and adding more matches to the match ID list using an iterative process. This will be done by using the new altered matchlist API call: [/api/lol/{region}/v2.2/matchlist/by-summoner/{summonerId}](https://developer.riotgames.com/api/methods#!/1033/3539)

<https://developer.riotgames.com/api/methods#!/1033>

This list of match IDs will then be used to query the in game match data in step 3.

**Step 3: In Game Data Extraction**

Once a match has been queried using the API call: [/api/lol/{region}/v2.2/match/{matchId}](https://developer.riotgames.com/api/methods#!/1027/3483) a JSON file will be returned containing all in game data for that specific match ID. This process of requesting match data must also be iterative. The JSON file returned will contain a lot of information that is not pertinent for analysis. This step will align with the feature selection step in the analysis approach. An example of match data and a list of all features included in the match data was included in the submission as a separate file.

**Step 4: JSON File Manipulation**

The returned JSON(Java Script Object Notation) file in its original returned state is too large and impractical for analysis. It contains many features that will not assist in the analytical process. All match stats files must be parsed by extracting necessary data points. Some points will contain continuous variables such as total gold, creep score, and turret kills. Other variables will return binary or Boolean statements. The JSON file manipulation step will be synchronous with the feature selection step from the analytical approach step. It may be prudent in this process to extract all potentially important data points including ones that may not initially seem relevant. The full list of data points that need to be extracted is listed in step 1 of the next section.

**Step 5: Exporting to Custom Dataframe Format**

Once all features for extraction are chosen they will be converted to CSV(Comma Separated Values) format. This is an important step that requires more refinement but this process will allow for the dataset to be analyzed using the R Language in RStudio. There are free to use JSON to CSV file conversion websites but they cannot be used for the project purposes. The final data frame will exist on a 2 dimensional format while the returned in game data is in a nested format, as is the nature of JSON files. This process must be completed in Python using a script that is also iterative for all matches. Tutorials on JSON file manipulation have already been researched and this step will most likely require some external assistance from someone with a higher level of expertise in the Python language.

**Step 6: Potential Application Deployment**

The completion of this step is dependent both on time constraints as well as the successful creation of a new model or metric. It is not my goal to create a functional service to users just for sake of creating an application. The model and application must provide tangible value to its potential user and I believe that this process can only be completed after a thorough analysis and testing of models.

**Step 1: Preliminary Feature Selection**

The initial list of features is listed in the Excel file Initial Feature Selection. This list will be reduced and curated based on the order of analysis steps as listed in next step. Many of the individual stats will be repeated 10 fold to reflect the number of participants within a game. I anticipate that this process will require more thought and diligence due to the fact that the selection of features will determine how the data frame is structured. It is very likely that new data frames will have to be created from the complete data frame in order to accomplish ad hoc analysis for various features. This will be done in RStudio using the dyplyr package and other data frame manipulation tools.

**Step 2: Correlation Analysis**

While the more commonly used Pearson correlation will be used in some cases for analysis, I anticipate that the Point-biserial correlation analysis will be much more frequently used. The dichotomous variable used the most will be win or loss result variable. Mean square contingency coefficient or Phi coefficient will be used for measure of association between two binary variables. This will be an important and revealing step for the entire project. In some ways, this step has the potential to be an arduous and time consuming task. However, these tasks will reveal new insights into the game and allow me to make mathematically proven statements about the nature of relationships within the game.

**Step 3: Evaluation of Selected Features**

This step is a continuation of the previous step. It is as much process and principle validation as anything to ensure the validity of any conclusions made using analytical techniques. I anticipate this step will require some guidance and more research but it should also be highly educational.

**Step 4: Analysis of Existing Performance Metrics**

Most if not all current metrics for individual performance evaluation such as gold difference, gold total, number of objectives, KDA and kill participation can be derived using the data that will be extracted. Many of these metrics are derived from simple formulas that can be programmed in R Language. All these features will be analysed using appropriate correlation, regression, classification and prediction techniques. The potential creation of a new model for performance evaluation exists in this step.

**Step 5: Recommendation and Creation of New Metrics**

This step exists in the event of a new model for performance evaluation being conceived in the previous step. This step may have been understood to be the crux of my project based on my abstract. However, after some contemplation and experimentation with various formulas using static match data, it was determined that my efforts will be primarily focused on the exploratory analysis. I do plan on continuing to design new models throughout the process and the focus on this step may change depending on how much the exploratory analysis reveals.

**Step 6: Model Veracity**

A potential model for performance measurement must provide valuable insight to the user. These insights should be able to give new information that a player can use to make assumptions or conclusions about their play with confidence. Measuring model veracity coincides with measuring the functional utility of the new model. A potential user must be able to answer yes to the question: Does this new metric reveal something about my performance that I otherwise could not have found out through existing methods of inference and observation?

**Conclusion**

I hope that I have achieved my goal of defining the overall conceptual theory of my project. I believe that the final submission will invariably look different than what you imagine this project to be but I have made my best efforts to define the process. I do understand that part of your position as an academic advisor is to restrain the student from veering off the task at hand by managing the scope of work to be done. It certainly would have been easier to use an existing dataset with less preparatory technical challenges. As I have stated in our previous meeting, it is important to me to complete a project of interest and passion. It is my assertion that undertaking a project of this scope will have more benefit to me academically and professionally; therefore I look forward to the challenges and breakthroughs.