data memo 1

Ellie Huang

2024-11-01

An overview of your dataset

What does it include?

It includes match variables: gameId, gameDuration, gameMode, platformId, queueId. It also includes gameId, gameDuration, gameMode, platformId, queueId, participantId, champExperience, champLevel, kills, deaths, assists, goldEarned, goldSpent, largestMultiKill, individualPosition, teamPosition, damageDealtToObjectives, damageDealtToTurrets, wardsPlaced, visionScore, teamId, win, killParticipation, controlWardsPlaced, goldPerMinute.

Where and how will you be obtaining it? Include the link and source.

From the Riot Games Developer Website – https://developer.riotgames.com https://developer.riotgames.com/apis#match-v5/GET_getMatchIdsByPUUID

About how many observations? How many predictors?

I think around 5000, querying match stats for various players, with around 25 predictors.

What types of variables will you be working with?

Numeric Variables: Continuous values that quantify aspects of player and team performance, such as gameDuration, champExperience, champLevel, kills, deaths, assists, goldEarned, goldSpent, largestMultiKill, damageDealtToObjectives, damageDealtToTurrets, visionScore, killParticipation, and goldPerMinute.

Categorical Variables: Discrete values representing categories or labels, which include gameMode, platformId, queueId, individualPosition, teamPosition, and win.

Is there any missing data? About how much? Do you have an idea for how to handle it?

To handle missing data: For numerical fields, I plan to use mean or median imputation. For categorical fields, I'll use the mode or potentially drop rows if key information (such as position) is missing and impacts model accuracy. For fields where missing values imply no action (like largestMultiKill), filling with zero is a straightforward approach.

An overview of your research question(s)

What variable(s) are you interested in predicting? What question(s) are you interested in answering?

I am interested in predicting player KDA. What is a player's KDA based on their in-game performance?

Name your response/outcome variable(s) and briefly describe it/them.

Outcome variable: KDA KDA in League of Legends reflects the kill and assist to death ratio. It is calculated by adding the total number of kills and assists and dividing it by the total number of deaths during the game.

Will these questions be best answered with a classification or regression approach?

These questions will be best answered with a non-linear regression approach.

Which predictors do you think will be especially useful?

I think match duration, champ experience, champ level, gold earned, gold spent, damage dealt, vision score, wards placed, individual position, team position will be especially useful.

Is the goal of your model descriptive, predictive, inferential, or a combination? Explain.

The goal of the model is primarily *predictive* with some descriptive elements. The focus is on accurately predicting a player's KDA based on match statistics, which involves using game data to make precise forecasts about in-game performance. However, the model will also provide descriptive insights into how various factors like experience, position, and vision control contribute to KDA, helping to understand key factors of player performance.

Your proposed project timeline

When do you plan on having your data set loaded, beginning your exploratory data analysis, etc?

Provide a general timeline for the rest of the quarter.

I plan to follow the same schedule as provided in the course syllabus.

Week 1: Pick topic
Week 2: Find data
Week 3: Tidy data
Week 4-5: EDA
Week 6-7: Run models
Week 8: Write-up
Week 9: Edit/review
Week 10: Final draft

Any questions or concerns

Are there any problems or difficult aspects of the project you anticipate?

Any specific questions you have for me/the instructional team?

I think the first difficult aspect of the project for me will be figuring out how to import the data from the API and cleaning the data, because it has a lot of extra factors that may not be significant. I also am wondering what type of model would work best, and how I can apply it.