Introduction to Data Science
Project Phase 2 - Report

Mohammad Yaser Azad | 810800003
Parham Javan | 810800008

Part 1: Preprocessing

1. Drop unnecessary columns (Columns like address of game which we don't need)
2. Handle Missing values (none found so we skip this part)
3. Data is stored in str so we need to convert them to list of Integers

Note: first team is represented as Blue(B) and second team as Red(R) and we aim to predict if blue team wins or lose

Part 2: Feature Engineering and Selection

1. Since one team status is meaningless, we calculate the difference of these two opposing teams (for example we calculate like Golddiff = BlueGold – GoldRed and after calculation we remove 2 initial columns)
2. We add new feature named like gold_last which is the status at last minute(30th minute) (note: we remove matches that are less than 30 and for rest we only examine these first 30 minutes)
3. Our data has data on each member of team since we have a lot data already and and sum of all these members are in the team we ignore these data)

Part 3: Dimensionality Reduction

1. We convert columns containing list to new columns of integer so we can analyze and calculate
2. Then we calculate PCA for 2 components variance is 56%
3. And for variance 95%+ we have to have 10 components and Explained variance with 10 components: 0.9584

Part 4: Evaluation Metric

Difrent Metrics

Accuracy: If your goal is to predict match outcomes (win/loss), accuracy is a straightforward metric to evaluate how often your model predicts the correct result.

Precision, Recall, F1-score: If your task involves predicting specific outcomes (e.g., predicting wins for a particular team), precision (ability of the model not to label a negative sample as positive), recall (ability of the model to find all positive samples), and F1-score (harmonic mean of precision and recall) are useful metrics.

Log Loss: For probabilistic predictions (e.g., predicting the probability of winning a match), log loss measures the performance of a classification model where the prediction input is a probability value.

Mean Absolute Error (MAE) or Mean Squared Error (MSE): If your goal involves regression tasks such as predicting game duration or player performance scores, MAE or MSE can measure the average magnitude of errors in predictions.

For our project we chose Accuracy , Why Accuracy?
Accuracy is a commonly used metric when dealing with binary classification tasks, such as predicting game outcomes (win/loss). It measures the proportion of correctly predicted outcomes among the total number of predictions made by the model. In the context of predicting League of Legends game outcomes(win/loss)

why not specific outcome is not important?
because one team wins and other loses so we have both and one doesnt have priority over the other one and if we calculate the diffrence the other way all results would reverse therefore Accuracy is chosen for main evaluation

Part 5: Model Training

Since the target variable is binary (win 1 or loss 0) the best model is Logistic Regression and as we can see has the highest accuracy among all models(the graphs has been plotted in the code)

Part 6: Feature Analysis (PCA)

We calculate and see the accuracy is lowered by 3-4% this is logical since now we significantly less features to analyze

Part 7: Overall Report and Discussions

Note: first team is represented as Blue(B) and second team as Red(R) and we aim to predict if blue team wins or lose

Note: we remove matches that are less than 30 and for rest we only examine these first 30 minutes after removal 6200 data remained of all 7300

Note: I have data normalized and saved but particularly didn't make a use of it

Challenges

-- its was much of a challenge to get use of data since first it was a str and after converting to list is was hard to manage since the number of components of each list is different unless we don't use all data and for example use the 30th minute only (for example gold stores gold for all the minutes as long as games continues)

--and although I converted first 30 values of all these list to new columns(simplest way but not best and most efficient one) we had problem handling them since it was around 210 columns and it is a lot and need lots of resourses