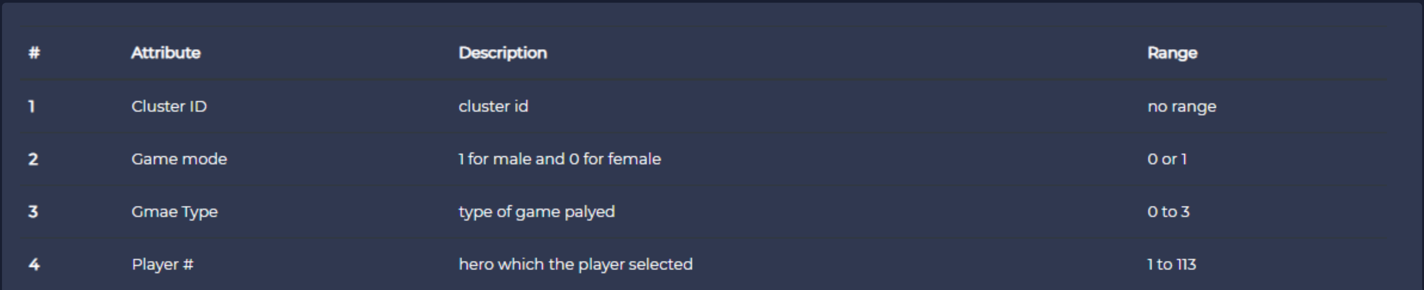
Exploratory data analysis

Data descriptive analysis

The training data has 92650 rows and 117 columns. The attributes or features is descripted In table 1. Game mode and game type are considered as caegorical nominal data and no ranks are given to either of them. Heros are also treated as nominal data. And no heros can be selected more than one player in a game. The condition is checked in the server using flask python, and error message is notified on such cases.



#	Attribute	Description	Range
1	Cluster ID	cluster id	no range
2	Game mode	1 for male and 0 for female	0 or 1
3	Gmae Type	type of game palyed	0 to 3
4	Player #	hero which the player selected	1 to 113

Table 1

The headmap showing the null valus in the dataset is shown in the fig1. There is 0% missing values is the training data matrix.

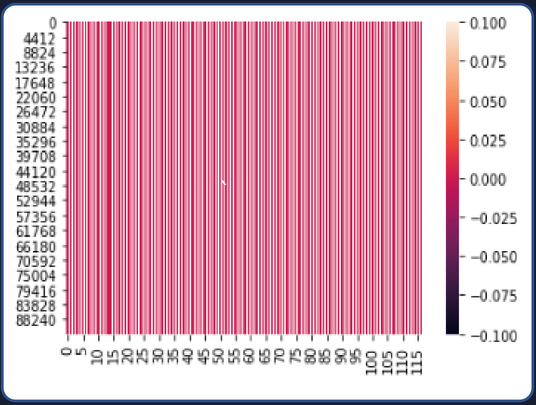


Figure1. heatmap depicting the null value in the data matrix

From the training data, we can see that both teams almost won the same number of games, in Figure 2 on the right.

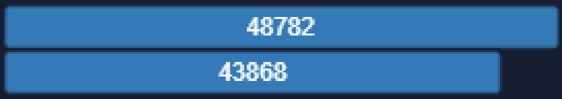


Figure 2:bar plot of number of games won by each team

On analysing Hero wise, we can see that all heroes are selected equally between both team members. The most selected hero is 8th(by numeric order), and is selected 32048 players(distributed as 16301 and15747 by each team.)

Correlation heat map is of the data is shown in the fig 2.

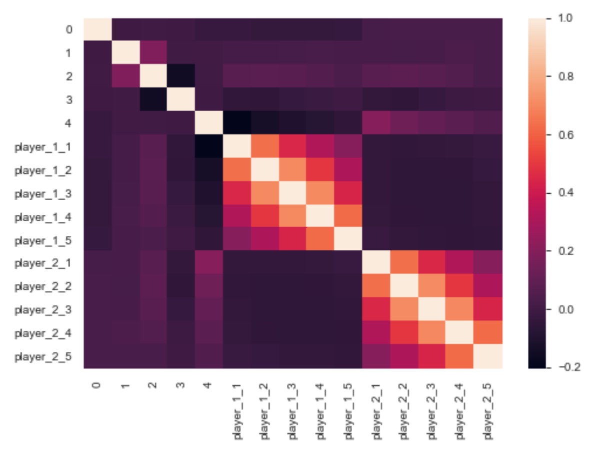


Figure2. Correlation heatmap showing the pearson correlation value

This the heat map of the data matrix after processing.I replaced the 113 columns of sparse data of heroes with 10 columns of players with the data of avatar selelcetd by them.

From the correlation map, we can see that there is no substantial correlation between the winning team and other features. But we can extract some correlation between the heroes selected within the team. SO there must be some pattern followed by the team members in selecting the team. Therefore simple regression models wont fit the data. An ensemble model of fitting the data is being used to build the model. A high number of weak learners would be needed to fit the data and optimize the loss function.

An xgboost Classifier is being used to train the model. Data is trained with low learning rate of 0.01 with a high number of estimators of 600. The model is trained using randomized search CV. And an accuracy of 60% was able to attain by the data despite of low correlation. It means that some heroes have substantially powerful than the fellow heroes.