

Predicting the Winner in CS:GO Matches

(STATS/CSE 780 course project)

Yan Min (Student number: 400546093)

2023-12-11

Abstract

This study explores the application of advanced machine learning techniques, specifically Random Forest and Deep Neural Network(DNN), in predicting the winners of CS:GO matches. We aim to use the past data of matches to forecast future match results. Leveraging the dataset provided by Skybox as part of their CS:GO AI Challenge, we preprocessed the data and then fed the data into the models. The performance of these two models are evaluated by accuracy, precision, and recall metrics. Our results yielded relatively high accuracies for both models, up to about 75%. The two machine learning models we developed improve the prediction and analysis of future match results, which is beneficial for the development of eSports.

Introduction

The dataset(“CS:GO Round Winner Classification” 2020) was originally published by Skybox as part of their CS:GO AI Challenge, running from Spring to Fall 2020. The dataset consists of around 700 demos from high level tournament play in 2019 and 2020. Warmup rounds and restarts have been filtered, and for the remaining live rounds a round snapshot have been recorded every 20 seconds until the round is decided. We want to use this dataset to predict the final winner, which can be seen as a classification problem.

There are totally 122410 entries and 97 colums in this dataset. The features have one boolean type, 94 float types and 2 object types. The number of entries is greater than the number of parameters. There are no null values in this dataset. We choose `round_winner` as the output value, which is consisted of the winners CT and T. There are other different types of features, such as `time_left`, `ct_score` and `ct_health`, which will be used as the input features. Figure 2 shows the distrubutions of selected features after feature slection.

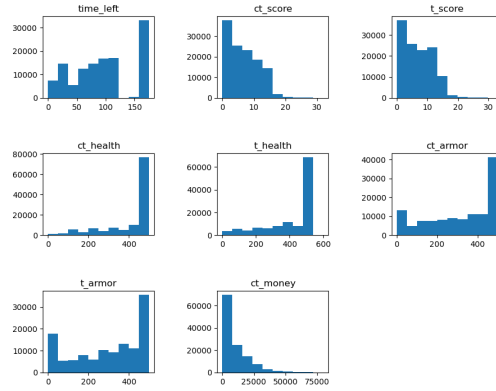


Figure 1: The distribution of selected variables after feature selection

Methods

Firstly, we plot the correlation matrix heatmap for the dataset to inspect the variables' mutual relationships. Secondly, we check whether there are missing values. Thirdly, we handle the outliers by substitute them into .95 quantile. Fourthly, we separate the original dataset into the training set and test set and the proportion is 0.7.

Then, we will use feature selection to find the best-suited predictors, since the number of variables is big, up to 95. We will select Forward stepwise selection (FSS) algorithm or hybrid approaches to select features. Then, we will use logistic regression and random forest classifier to build the models. During the modeling process, there may exist circumstances in which we need to do parameter tuning. We will change the size of the training dataset to find the best prediction result. Since the results of prediction is binary, either bankrupt or not, it's easy to interpret the predictive results. We will use the ROC curve to find the best cutoff.

After training the models, we will use the testing set to make prediction and then compare to the real values. We will caculate the accuracies for different models and compare them.

Results

There are 6819 instances and 96 features in this dataset. The fist one, **Bankrupt?**, is categorical, either 1 or 0. The other 95 variables are numeric. Figure 2 shows how these variables distribute. Some of them are left skewed and I will standardize them.

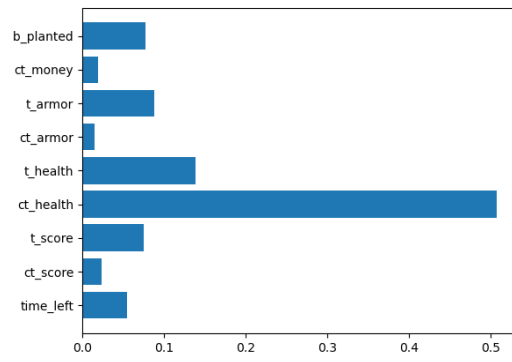


Figure 2: The distribution of all the variables

Conclusion

Supplementary material

References

“CS:GO Round Winner Classification.” 2020. <https://www.kaggle.com/datasets/christianlillelund/csgo-round-winner-classification>.