CS:GO Round Winner Prediction

1st Mustafa Mert Sandal
Computer Engineering
TOBB ETU
Ankara, Turkey
191101030

2nd Murat Gencer
Computer Engineering
TOBB ETU
Ankara, Turkey
191101043

***Abstract*—Today, online games are developing with the increasing interest in e-sports. CS-Go is one of these games. Cs-Go is an online game that is known worldwide and tournaments are held around the world on this game. Since it is so popular, it is easy to collect data and it organizes machine learning challenge for game result. This project aims to train a model for CS-GO that predicts rounds for the game.**

I. Introduction

Cs-Go is one of the most popular computer games today. This game features many professional teams and esports players. Tournaments are held around the world every year and it has a large number of followers and investors. The Project's goal is to create ML model that gives good prediction for CS-GO.

A. Data

The dataset was originally published as part of Skybox. CS:GO AI Challenge from Spring to Fall 2020. The dataset consists of 700 demos from the top tournament. Play in 2019 and 2020. Warm-up laps and restarts filtered and a tour for the remaining live tours snapshot saved every 20 seconds up to tour the decision is made. After initial publication, it has been pre processed and flattened to increase readability and make it more convenient. Total number of snapshots is 122411. The dataset also has 97 columns. In this project, data shared by Skybox was used and a model was tried to be trained. The model was created to predict who would win the round with the new data entered.

II. Related Works

A. Kaggle Notebooks

In Kaggle there are 25 notebook about the topic that uses the dataset was originally published by Skybox as part of their CS:GO AI Challenge, running from Spring to Fall 2020. Some of the notebooks make just exploratory data analysis(EDA). Also in Kaggle there are different notebooks. The notebooks use different dataset.

B. Articles

On CS-Go prediction published articles by universities. An Article that was published by Chalmers University of Technology uses CS-GO competition's data. The data collected from FACEIT website. FACEIT is an electronic sports platform. Collecting data from website, used the Selenium chrome API extension for C. The model clustered features. For this used k-means Clustering. The model uses a feedforward neural network algorithm.

The another article was published by Czech Technical University in Prague. The article tests non-neural network model and different designs of neural networks model. There are three non-neural network model. These is Logistic Regression, Random Forest Classifer, K-Nearst Neighborhood Classifier. Also, the model uses linear models. Linear model is instance of a class that inherits from PyTroch's nn.Module. Another model is convolutional model. Convolutional model is an instance of a class, inheriting from nn.Module. For the article data collects from Kaggle. The dataset include CS-GO Professional matches. The dataset is different from the dataset used in the Project

C. Paper

There is Allen Rubin's Predicting Round and Game Winners in CSGO paper. This paper used 1229 professional game rounds as dataset. There are 17 attributes and 2 results in the dataset. In this paper, tree-based classifiers and a neural network is being used. For XGBoost applications, scikit-learn implementations of random forest and multilayer perceptron classifiers are used together with the xgboost package. A 70/30 train/test split is used. Most of the time, they can predict tour and match winners correctly. For the round level, the Random Forest model performs better in every metric, while for the gaming level, the XGBoost model performs better in every metric.

The project is being used principal component analysis(PCA) for feature selection. It uses 4 models used in notebooks and the article to create a model, and unlike the studies that have been done, it uses the TabNet algorithm. TabNet is a good model that is based Deep Neural Network for tabular data.

III. Methods

There are too many features in the dataset, so preprocessing is required on the data. The data that will not contribute to the model training are determined first. Visualization is done in this step. This visualization helps to better understand the data.

A. Preprocessing

During this visualization, it is observed that 2 features are object type and 1 feature is bool type. In order to use these 3 features in training, encoding is done using the label encoder.

During preprocessing, For a good training, it is necessary to control the copies. Data scaling is required to train the model, if the numeric value is not scaled, the numeric values can be very different from each other, making it impossible to train the model. In order to improve model training, data with high correlations are removed. As a result of these processes, the data becomes suitable for processing, and then feature reduction is performed with principal component analysis (PCA). The PCA process shows that we can train a good model with 68 features without making much changes on the data. According to PCA, we can maintain 95% variance with 68 features.

B. XGBoost

The XGBoost algorithm is used as the first model. XGBoost is a high-performance version of the Gradient Boosting algorithm optimized with various modifications. XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems. Software and hardware optimization techniques have been applied to obtain superior results using less resources. It is cited as the best of the decision tree-based algorithms.

C. Logistic Regression

Then the logistic regression algorithm is used. Logistic regression is defined as a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. Logistic regression estimates the probability of an event occurring based on a given set of independent variable data. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transform is applied to the probabilities. Logistic regression is an important technique in the field of artificial intelligence and machine learning.

D. Random Forest

Third, the random forest algorithm is used. Random forest algorithm creates many trees. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It randomly selects certain features from the training data and it is trained with random samples. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

E. Deep Neural Network

The deep neural network is trained as the next model. The main purpose of a neural network is to receive a set of inputs, perform progressively complex calculations on them, and give output to solve real world problems like classification. It often gives inefficient results for tabular data compared to other algorithms. However, since TabNet, which makes neural networks better for tabular data, will be used in future models, it is also desirable to learn the performance of neural networks.

F. TabNet

Finally, the actual targeted TabNet model is trained. this model. TabNet is a deep tabular data learning architecture that uses sequential attention to choose which features to reason from at each decision step. The TabNet encoder is composed of a feature transformer, an attentive transformer and feature masking. In this project, this algorithm is used to see how much it has improved the neural network algorithm and to compare it with other models.

IV. Experimental Results

Models are trained with the algorithms given in the previous section and the results of these models are examined.

A. XGBoost

For the XGboost algorithm, the number of estimators and the maximum depth parameters are tested with a few values, and as a result, the model is trained with these parameters and the test result is obtained. Before optimizing the parameters, the default value is 100, 300, 500 as the number of estimators, and the default values of 10, 30 and 50 for maximum depth are tried. As the optimized parameter, the number of estimators is 500 and the maximum depth is 50. If the test data is estimated with these parameters, the results are accuracy score 79% for the parameter(default , default) accuracy score for parameter(10 , 100) 84% accuracy score for parameter(30 , 300) 85% accuracy score 86% for parameter(50 , 500). Thus, the optimized parameters become 50 to 500.

B. Logistic Regression

The algorithm implemented in the project are logistic regression, Random Forest, Deep Neural Network, XGBBoost and Tabnet. In the Logistic Regression algorithm, it is tried with more than one parameter to find the parameter that will give the most accurate result. These parameters are penalty function, optimization parameter and regularization parameter. Values 'l1', 'l2' for Penalty parameter 'newton-cg', 'lbfgs', 'liblinear' values for the optimization parameter and different numeric values for regularization are tried. Among these values, the most optimized values are 'l2', 'newton-cg' '100'. With these values, the following results are obtained.

[accuracy] : 0.74710 (+/- 0.00434) [precision] : 0.75256 (+/- 0.00519) [recall] : 0.75020 (+/- 0.00375) [f1] : 0.75137 (+/- 0.00394) [test accuracy] : 0.74463

C. Random Forest

An optimized result is obtained by changing the 'criterion' , 'max_depth' , 'n_estimators' parameters with the Random Forest algorithm. It may be necessary to wait for a long time to optimize these parameters. Finding optimized parameters is not easy for the system. This project has criterion = 'gini' , max_depth = 80 , n_estimators=500 as optimized parameter. When the test data is tested with these parameters, the accuracy score is 86%.

D. Deep Neural Network

In the deep neural network(DNN) algorithm optimization is made according to the number of dense and the number of neurons found in these dense, also activaiton function is an important parameter for optimizing DNN. In this project, 5 Dense is used as the number of Dense. These Dense contain 200, 100, 50, 25,1 neurons, respectively, and the activation function is chosen as relu because this project is a classification project. As a result of these parameter selections, the deep neural network model gives a low accuracy score. This is expected because it is known that Deep Neural network methods are not very efficient on tabular data. Tree-based structures are better in tabular data. It is also seen in this project. The prediction scores of the random forest and XGboost algorithms give a more successful score.

E. TabNet

Another algorithm used in this project is TabNet. TabNet is a deep neural network based algorithm. Deep neural network-based algorithms actually do not work well with tabular data. But TabNet is a DNN-based algorithm developed for tabular data. Especially TabNet is used in this project. The aim is to show that TabNet can give good results on tabular data. For TabNet optimization, changes are made on the learning rate epoch number and patience parameters. At the end of these optimizations, having the epoch number of 500, patience 50 and the learnin rate of 0.5 gives us a good result. Also, mask_type and scheduler_params parameters are set for TabNet. $mask\_type =' entmax' andscheduler\_params = \{"step_size" : 10, "gamma" : 0.9\}$. This model gives the accuracy score as 77% as a prediction after optimization.

V. Conclusion

All in all, csgo is a popular game played by many youngsters. It brings a lot of data along with its popularity. The aim of the project is to know which team will win the round for this game. In this direction, the data is first visualized. As a result of this visualization, the necessary pre-processes are decided. These necessary pre-processes are done and the data is made suitable and comfortable for machine learning.

Then, familiar artificial intelligence algorithms are used for the database we have. In addition, deep neural network and TabNet algorithm are used for this data set, which has few or no examples. These two algorithms are specially selected. It is desired to show that the neural network algorithm gives bad results in tabular data. It also shows how to improve the neural network using TabNet. In the end, XGBoost and Random Forest give the best results. Next comes TabNet, then Logistic Regression, and lastly Neural Network.

Finally, the little used TabNet algorithm for the dataset is shown. It is believed to improve the neural network for tabular data.

References

”TabNet: Attentive Interpretable Tabular Learning,
Allen Rubin’s Predicting Round and Game Winners in CSGO:
https://osf.io/u9j5g/
https://www.kaggle.com/code/christianlillelund/predict-winners-in-cs-go-with-keras-80/notebook
https://www.kaggle.com/code/sbezvitniy/cs-go-winner-prediction
https://github.com/jefersonmsantos/counter_strike/blob/master/machine_lea
e%20Learning%20prediction%20model.ipynb
http://publications.lib.chalmers.se/records/fulltext/256129/256129.pdf
https://dspace.cvut.cz/bitstream/handle/10467/99181/F3-BP-2022-Svec-Ondrej.predicting_csgo_outcomes_with_machine_learning.pd
https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/
https://paperswithcode.com/method/tabnet: :text=TabNet%20is%20a%20de