Modeling of Counter Strike: Global Offensive Rounds using Machine Learning

A PROJECT

Submitted by

Devanaboina Rohit

Reg. No. 1920181

Under the guidance of

Prof. Deepak Joy Mampilly

Asst. Professor, School of Business and Management

In Partial Fulfillment of the Requirements for the Award of the Degree of

BACHELOR OF BUSINESS ADMINISTRATION

CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

SCHOOL OF BUSINESS AND MANAGEMENT
CHRIST (Deemed to be University)
BANGALURU
2022

CERTIFICATE

This is to certify that the project submitted by Devanaboina Rohit (Reg.No:1920181) titled “Statistical Analysis and Modeling of Counter Strike: Global Offensive Rounds” submitted to CHRIST (Deemed to be University), in partial fulfillment of the requirements for the award of the Degree of Bachelor of Business Administration, is a record of original study undertaken by Devanaboina Rohit, during the period 2021 – 2022 in the School of Business and Management at CHRIST (Deemed to be University), Bangalore, under my supervision and guidance. The project has not formed the basis for award of any Degree / Diploma / Associate ship / Fellowship or other similar title of recognition to any other University.

Place: Bengaluru	Prof Deepak Joy Mampilly
Date: 21/04/2022	Asst. Professor, School of Business and Management

Dr Amalanathan. S

Head - School of Business and Management

DECLARATION

I, Devanaboina Rohit, hereby declare that the project, titled “Statistical Analysis and Modeling of Counter Strike: Global Offensive Rounds”, submitted to CHRIST (Deemed to be University), in partial fulfilment of the requirements for the award of the Degree of Bachelor of Business Administration is a record of original and independent study undertaken by me during 2021–2022 under the supervision and guidance of Prof. Deepak Joy Mampilly, School of Business and Management. I also declare that this dissertation has not been submitted for the award of any degree, diploma, associateship, fellowship or other title to any other Institution/University.

Place: Bengaluru

Date: 21/04/2022 Devanaboina Rohit (1920181)

ACKNOWLEDGEMENT

I, Devanaboina Rohit would like to express my profound gratitude to all those who have been instrumental in the preparation of this research Project Report. I wish to place on records, our deep gratitude to my project guide, Prof. Deepak Joy Mampilly, for guiding me through this project with valuable and timely advice. I would like to thank Dr. (Fr). Abraham V M, Vice Chancellor, Dr. Jain Mathew, Dean and Dr. Amalanathan S, Head-School of Business and Management for their encouragement.

Last but not the least; I would like to thank my parents and friends for their constant help and support.

Devanaboina Rohit

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Introduction

Since ancient times, people have sought to predict the future. The Mesopotamians made primitive attempts to predict Lunar Eclipses as early as the mid-seventh century BC [1]. Modern prediction and forecasting techniques make use of Statistical, Machine Learning and Data Mining. This field of study is known as Predictive Analytics. [2]

Sports Prediction is one application of predictive analytics. Its objective is to accurately predict the proceedings and/or outcomes of a sporting event. Modern Sports Prediction relies on statistical modeling of sporting events. [3]

Predictive Analytics have always been a popular subject in the field of sports analytics, beginning with classical statistical analysis and modeling and continually developing in sophistication through innovations such as modern-day Machine Learning models & geospatial tracking. However, literature on **eSports predictive analytics** is much leaner compared to its traditional counterparts.

With experts forecasting that **eSports viewership will at least match**, if not exceed, that of traditional sports leagues such as the **NBA and MLB** [4], it is important to bolster the existing capacity and capabilities of Sports predictive analytics for the sake of viewers, analysts, players and researchers. The objective of this study is to do exactly that for the popular eSport Counter Strike: Global Offensive

1.2 eSports

eSports, or electronic Sports, is as a form of sports where the primary aspects of the sport are facilitated by electronic systems - the input of players and teams as well as the output of the eSports system are mediated by human-computer interfaces. [5]

The term “eSports” dates back to the late nineties. One of the earliest reliable sources that use the term “eSports” is a 1999 press release on the launch of the Online Gamers Association (OGA) in which eSports were compared to traditional sports. [7]

eSports has grown to join the ranks of traditional sports, with the estimated number of eSports fans numbering 474 million worldwide as of 2021, a ~20% increase from 2019. Forecasts predict this will rise to 576 million by 2024. [6]

It has been noted that there are two schools of thought when it comes to eSports Research: [7]

i. eSports Research is to be treated in the same manner as traditional Sports Research
ii. eSports Research is to be treated as a field separate from traditional Sports Research

1.3 Counter Strike: Global Offensive

Counter-Strike: Global Offensive (CS:GO) is a competitive First-Person Shooter (FPS) video game developed by Valve and released for PC in 2012. It is one of the largest multiplayer games by user base, with over 500,000 average concurrent players in September, 2021. It is also the largest eSport in terms of prize pool in 2021.

1.4 Need for the Study

Predictive analytics for eSports have their origin in traditional sports analytics, with many of the same techniques and models being adapted to the eSport model, with mixed results. The gaps in advanced eSports analytics are evident - the variety and accuracy of current Player Rating, Match prediction and Geospatial Analysis models (amongst others) in eSports pale in comparison to those of traditional sports.

As eSports are played on digital platforms, where capturing large volumes of data with exceedingly high granularity is a relatively low-effort task, the data available for analysts is orders of magnitude above that of traditional sports. Users are able to capture game, player and environment/context data with granularity down to the millisecond. The physical and technological limitations of data collection tools and techniques used in traditional sports make achieving similar levels of data granularity prohibitively expensive or even impossible.

Researchers with an "eSports-first" approach have however been successful in developing eSports-specific analytical frameworks built from the ground up, many of which are better at taking into account the unique context eSports operate under and better capture the richness of information provided by granular eSports data vis-a-vis traditional sports analytics models. In essence, they succeeded in better capturing and factoring in those elements that differentiate eSports from their conventional counterparts.

My objective is to apply proven statistical and ML techniques on CS:GO data, in order to more effectively model CS:GO competitive matches and the context surrounding them. My goal is to improve the accuracy and other key metrics of CS:GO match outcome predictive models.

The opportunity to conduct data analytics on highly structured events as provided by eSports in general, and CS:GO in particular, has the potential to accelerate research in predictive analytics as a whole, with finding and implications extending to traditional sports analytics or even more high-priority fields of analytics if the opportunity is sufficiently exploited.

CHAPTER 2:
REVIEW OF LITERATURE

2.0 Articles Reviewed = 20

2.1 Why Sports Prediction is Difficult

[9] (Aoki et al., 2017)

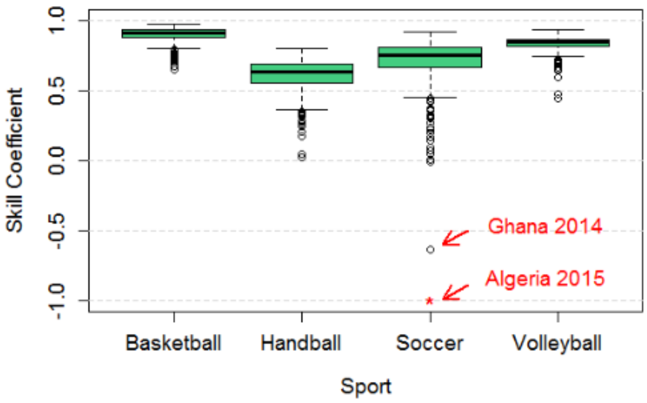
In every sport, there exist two factors that determine the outcome of a match: **SKILL** and **LUCK.** Sports outcomes are based on a mix of skills and good and bad luck.

- Skill refers to those factors player/teams can control and which influence their chance of success in a match– players' technical prowess & teams' strategy are examples of this.
- Luck refers to those factors that contribute to the outcome of a game but are cannot be effectively influenced by players – the characteristics of a Cricket pitch, the location of the match (Home/Away) etc. Luck also includes a number of factors that are either **unknown** or **too complicated to anticipate**, such as the exact trajectory of a sports ball (which has a level of unpredictability to it)

A study conducted by Aoki et al. (2017) provides a clear demonstration of this principle.

Matches in several major sports were modeled using factors that reflected skill in their respective sports. These "Skill-based" models were then used to predict the outcome in their respective sports. These predictions were then compared to the actual results, with the deviations between the two indicating **the level of "luck"** involved in each sport.

It was found that all sports tested had a significant luck factor involved in the outcome of matches. Luck had a smaller influence in some sports such as Basketball, which had a high Skill Coefficient, whereas luck had a larger role in certain sports such as Handball and Soccer. But nonetheless, luck had a say in the outcome in the matches of all the major sports tested.



Thus it is clear that Luck & Skill both have an impact on match outcomes. The relevance of this characteristic of sports is this –luck exists in sports, and this luck factor is very difficult to model and predict.

<u>LEARNING OUTCOME</u>

That is why Sports Prediction is difficult – Luck is difficult to model and predict.

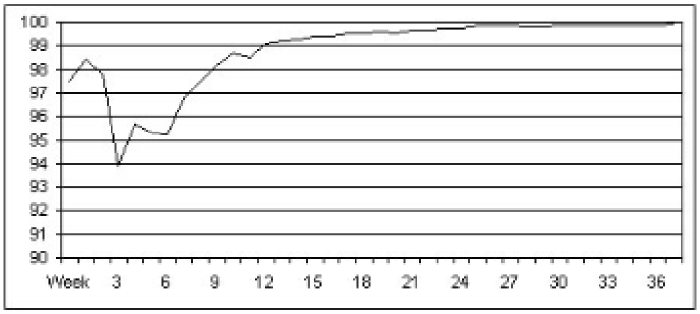
In a way, luck limits how accurate sports prediction models can be. This barrier cannot be crossed unless the factors falling under Luck can be measured and modeled – a task that although is becoming more feasible with advances in techniques, is still a difficult problem to solve.

2.2 Sports Prediction

[18] (McCabe and Trevathan, 2008)

In 2008, an attempt was made to test the power of AI in predicting the outcome of various sporting events, including Association Football. The study found that it is possible to create AI-based sports prediction models that have greater predictive power than that of human experts.

The general AI sports prediction model developed by them won 1st place at the major international tipping (Prediction) competition called TopTipper, beating thousands of (human) contestants of varying skill levels, from year to year. The chart below shows what percentage of participants the AI Model beat in terms of predictive accuracy, on a weekly basis in the TopTipper competition of 2006-7.



The AI Predictor's Performance in TopTipper Competition (McCabe and Trevathan, 2008)

[19] (Min et al, 2007)

Researchers undertook this study to create a model to predict Football match outcomes with the hypothesis that a holistic, statistical model can be created that can maximize the information derived from limited football data (qualitative and quantitative) and provide accurate match predictions.

The result of this study was the creation of FRES (Football Result Expert System). This framework divides a single football match into ten time frames to apply in-game time-series approach. And in each frame of a match, an attempt is made to model the reasoning that might be carried out by a good head coach. Both teams infer the current state of the match and derive corresponding strategies. This can be viewed as a knowledge-based in-game time-series approach; using it enables FRES to give realistic, and somewhat more accurate, predictions

FRES predicted six countries out of the actual top eight countries of the World Cup 2002, while the historic predictor and the discounted historic predictor predict five – the three predictors are relatively close in performance.

[3] (Claudino et al, 2019)
In this study, researchers conducted a systematic review of AI techniques used for Injury Risk Assessment and Performance Prediction in Team Sports

It was found that Artificial Neural Networks were the most widely used in Injury Risk assessment, with Decision Trees and Support Vector Machines second and third place respectively.

In Performance Prediction, ANNs topped the rankings again. However, the results were more even in this field with Decision Trees, Markov Process and Support Vector Machines showing significant use.

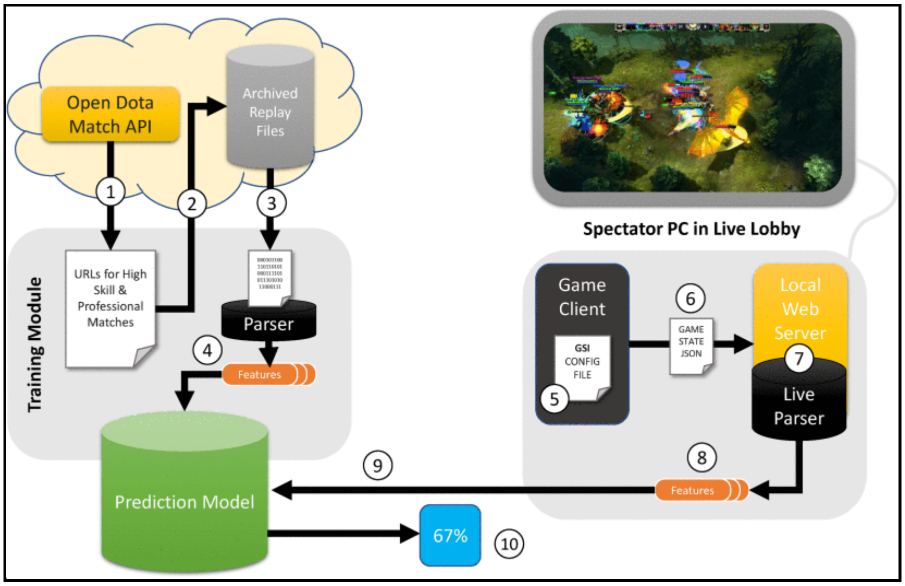
2.3 eSports Prediction

[10] (Makarov et al., 2018)
In this study, the TrueSkill Elo Rating system, a system similar to that used in rating chess players, was used to rate players & teams in the DOTA 2 and CS:GO (two popular eSports). The hypothesis of this study was that the outcomes of these two team-based Esports can be determined using systems applied traditional sports rating and prediction.
The TrueSkill rating of each player and team was calculated, and the team with the highest rating was predicted winner. The results were compared to a random selection for benchmarking. The **TrueSkill rating system is a better than random method of predicting the winner** of any particular match-up.

<u>LEARNING OUTCOME</u>: **Traditional sports prediction methods can be successfully applied to eSports.**

[11] (Hodge et al., 2019)

In this study, researchers attempted to create a Model that could predict the outcome of DOTA 2 matches (a popular eSport). The objective of the study was to build a model that could predict the outcome based on just the **first 5 minutes of the match** using real-time data.



The Prediction Workflow w/ Real-Time Data (Hodge et al., 2019)

Given above is the workflow used by the researchers to collect data, train the models and generate predictions using real-time data.

The hypothesis of this study was that player' actions before and during the first 5 minutes of a DOTA 2 match is sufficient data to predict the outcome of the match.

Various ML algorithms were used to process the independent data in order to predict the outcome of a match within the first 5 minute of the match going live. The data was collected and analyzed live, in game using an API based prediction program. The **predictive power of the models tested ranged from 70-75 %,** significantly better than a random guess.

LEARNING OUTCOME: It is possible to build effective predictive models that require only **minimal amounts of data** and **that can process real-time data** and output predictions in a **timely manner.**

[12] (Wang, 2019)

In this study, the researcher attempted to create a model that could predict the outcome of League of Legends (LOL) games based on game, round and player statistics.

The hypothesis for this paper is that the outcome of LOL games can be predicted using Logistic Regression and Decision Trees, commonly used ML models.

Various features were fed into the aforementioned models in WEKA. They were allowed to modify the weights for each of the features, optimizing predictive power. Though the model has decent performance, **the task at hand is too difficult** for standard models to fully learn. Due to the complex nature of team dynamics and interrelatedness of metrics, standard models cannot fully capture the game in its entirety. Modifications to capture the team dynamics of the game and eliminate redundancies can improve the performance of the model

LEARNING OUTCOME: Traditional Machine Learning methods may not be able to effectively model highly complex events, such as the ones often seen in some eSports.

[5] (Hamari and Sjoblom, 2017)

In this study, the researchers delve into the reasons why people watch eSports. In order to understand the motivations behind eSports consumption, the researchers made use of the Motivation Scale for Sports Consumption (MSSC), a widely adopted measurement scale for studying the reasons why people consume sports content.

Factors measured in the MSCC include Vicarious achievement (Empathizing and co-living the achievements of teams and players), Aesthetics (The appreciation of the beauty and gracefulness inherent in the sport) and Drama (The enjoyment of the drama, uncertainty and dramatic turns of events) amongst other factors.

The results indicated that escaping everyday life, acquiring knowledge from eSports, novelty and the enjoyment of aggression were positively and statistically significantly associated with the frequency of watching eSports.

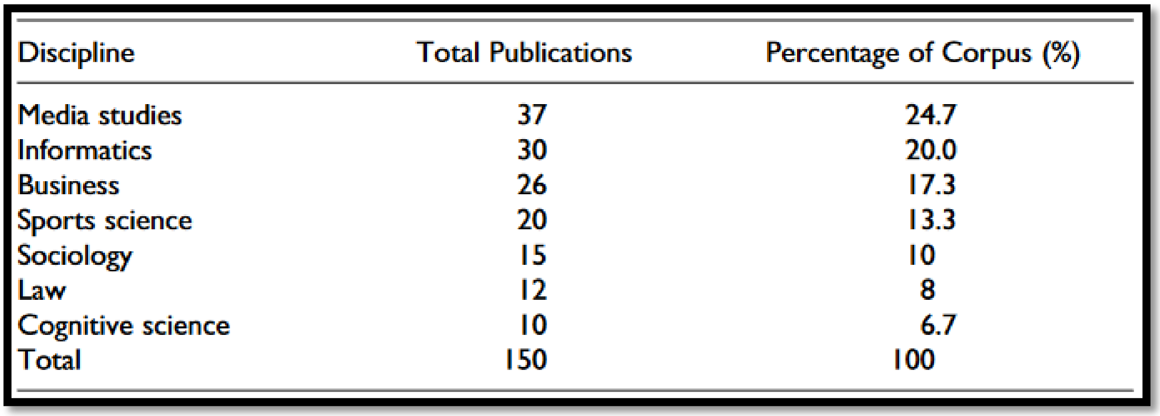
[7] (Wagner, 2006)

In this study, the author goes into detail about eSports as a Scientific Field of Study. The author begins by defining the term eSports, labeling it as physical and mental sporting activity conducted though ICT technology.

The author points out that eSports research could lead to insights into other fields. For example, the study of high-performance eSports teams can help further research into this topic in the domain of Management.

[8] (Reitman et al, 2019)

This study analyzed the distribution of eSports Research Papers across related fields from 2002 to 2018 revealed that **Media Studies** and **Informatics** were the most common Disciplines in which such eSports papers were published, with Business and Sports Science following closely behind. These four disciplines received 70% of eSports research over the nearly 2 decade period.



Discipline	Total Publications	Percentage of Corpus (%)
Media studies	37	24.7
Informatics	30	20.0
Business	26	17.3
Sports science	20	13.3
Sociology	15	10
Law	12	8
Cognitive science	10	6.7
Total	150	100

[20] (Melentev et al, 2020)

In this unique study that crosses the digital boundary, researchers attempt to We use EEG recording to monitor the players' performance status and use machine learning to find the correlation between EEG recordings and eSports athletes 'performance.

A Mitsar EEG Monitor & the Mitsar Data Studio was used to collect and process EEG data, while Python was used to conduct the modeling using ML models.

One set of models were trained to predict whether a given player was a casual or professional player based on EEG Readings. The Models preformed well, with the Gradient Boosting + Grid Search model achieving an accuracy of 92% and an F1-score of 95%.
Another set of models were trained to predict whether a given player was tired or not based on EEG. These models also performed well, with the GB + GS model achieving an 88% accuracy rate.

[21] (Ghazali et al, 2021)
In this study, researchers attempted to predict the placement/rankings of players in a game of PlayerUnknown's Battlegrounds using machine learning techniques.

A public PUBG game statistics database was used, comprising 29 variables related to the game and players. 3 models were tested – Regression Tree, Linear Regression and Support Vector Machines. The results indicated that the support vector machine (SVM) has the highest performance and better prediction model than the regression model and linear regression model. However, it is faster to train the model using a decision tree model and it is also easier to interpret the model.

[22] (Yang et al, 2022)

In this study, researchers attempted to create a real-time win prediction model for the game Honor of Kings. This is very similar to the objective of my research project, with the main difference being the game being modeled.

The researchers collected anonymized game data directly from the game publishers for this study. They used a Deep Learning model called Two-Stage Spatial-Temporal Network for the task of continuously predicting the win probability in a round of Honor of Kings. They used this model due to its interpretability and ability to analyze & output predictions in real-time. The peak accuracy of this model is 78.5%, when 10 minutes have passed since the start of the round.

2.4 Counter Strike: Global Offensive

[13] (Björklund et al, 2018)

In this study, researchers attempt to determine the influence of team composition in the outcome of CS:GO matches.

Hypothesis: The outcome of CS:GO Games is significantly influenced by the composition of the team. The 5 players on each side in any given CS:GO match and their role within their team has a significant effect on match outcomes**.**

Methodology: Around 50 features - all related to the game, rounds and players – were used to build the predictive model. In additional to these variables, the **specific role** each player played in their team (Entry, Support etc.) were determined by applying **Clustering Algorithms** to categorize players into the various role. Based on this data, Neural Networks were built to predict the outcome of a given match.

The networks that used team composition data could predict the games with an accuracy of 65.11%. Networks that rely only on win rate attained prediction accuracy of 58.97%.

LEARNING OUTCOME: Team composition has significant effect on the outcome of CS:GO matches (which also mean it can provide significant predictive power to models that use this data).

[24] (Rubin, 2022)

In this study, the researcher tested various models on the basis of their performance in predicting the outcome of CS: GO matches.

Hypothesis: Machine Learning and AI models can effectively predict the outcome of CS:GO matches.

Methodology: 45 CS:GO matches that took place in 2021 were used to collect the data on which the models were to be trained. 15 attributes repeated across 10 players were the features used in this study, along with the current score of each team.

3 models were tested – Random Forest, XGBoost and Multilayer Perceptron (MLP). Random Forest performed best in predicting round winners with a 64% accuracy rate, while XGBoost performed best in predicting the overall winner of a match with a 62.37% accuracy rate. Both of these are better than the benchmark accuracy rates.

LEARNING OUTCOME: Machine Learning Models can be effective in predicting the winners of CS:GO rounds and matches

[14] (Švec, 2022)

In this study, a comprehensive study of various methods and models was conducted in order to identify the best ways of predicting the outcome of CS:GO matches.

Methodology: Match Data was scraped from HLTV.com, a reputed source for CS:GO data. 3 ML models and 3 Neural Network models were tested in this study. Logistic Regression, Random

Forest and K-Nearest Neighbors are the ML models. Linear, Convolutional and Embedding models are the Neural Network models.

Random Forest was the best performing ML model with a Test accuracy of 63%, whereas the Convolutional Neural Network was the bests Neural Network model with a 59.8% accuracy. However, an Elo based model akin to the ones used in traditional sports prediction outperformed all these models with an accuracy rate of 64%.

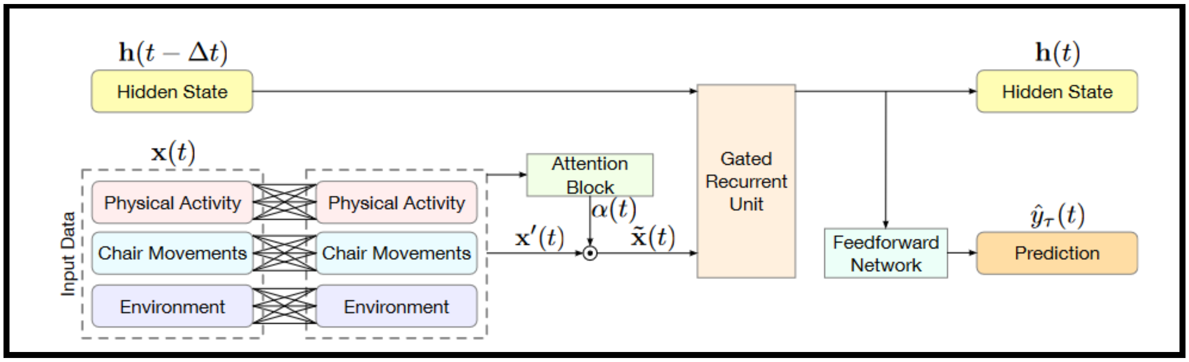
LEARNING OUTCOME: Traditional Sports Prediction Models have the potential to outperform even the most advanced ML and AI models.

[15] (Smerdov et al, 2020)

This study takes a unique approach to eSports prediction, relying on real-world data to predict players' performance in the virtual arena.

Hypothesis: Real-world data such as players' physiological signs, chair movements and environment conditions can be used to predict player performance in game (specifically CS:GO in this case).

Methodology:



The Recurrent Neural Network Architecture used in the Study (Smerdov et al, 2020)

Using sensors, researchers collected data on players' chair movements, physiological signs and environment conditions. They then used Recurrent Neural Network (RNN) and RNN with attention - along with non-AI models - to predict players' performance in Counter Strike: Global Offensive, a popular eSport.

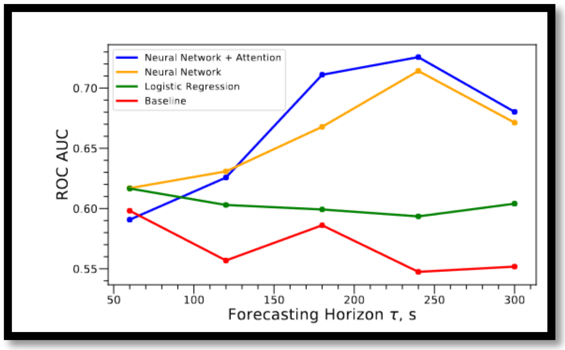
The results show that both AI models outperformed their ML counterparts in this task. The RNN with attention performed the best across all time horizons, indicating that the attention mechanism provides the RNN with a significant improvement to predictive power.

LEARNING OUTCOMES

1. eSports Prediction does not have to be restricted to in-game data. The real world can provide useful insights as well
2. AI models tend to outperform ML models when dealing with large amounts of complex data

[16] (Xenopoulos et al, 2020)

In this study, researchers attempted to determine the value of the various actions taken by players in CS: GO.



Various player metrics were jointly analyzed using ML techniques alongside round by round data (each game consists of around 30 rounds). Based on how the former affect the outcome of the latter, the model learns what outcomes are important, including the context which makes them important. Based on this learning, the model can take player statistics and round statistics and create an intelligent rating for each player in a given map Thus, a context-aware framework to value players' actions based on changes in their team's chances of winning was developed.

The rating created is stable, discriminant & independent - making it a suitable rating system that provides valuable information about player performance.

LEARNING OUTCOME: Machine Learning isn't restricted to predicting the outcome of matches. It can be used to derive previously unseen insights into CS:GO and its complexities.

[17] (Petri et al, 2021)

In this study, researchers attempted to use Reinforcement Learning models to predict the Map selections made by teams.

In every competitive game of CS:GO, teams vote on which maps are to be played in that match. There is a strategic element to this decision, and being able to predict the opposing teams' picks ahead of time can give teams a significant advantage.

According to the researchers, The Pick/Ban policies of teams are inefficient, with significant room for improvement that can lead to improved performance. Contextual Bandits (statistical models based on reinforcement learning) were applied to the Map Pick/Bans of various CS:GO matches. These optimal actions were compared to actual Pick/Ban actions of teams, to identify deviations from the statistically optimal actions.

The results indicate that teams using the chosen policy instead of their traditional map-picking process can increase their expected win probability by 9 to 11 percentage points, depending on the policy used. This is a substantial advantage for a best-of-3 match, since the model could confer that added win probability to all three map choices.

LEARNING OUTCOME: Advanced analytics and models could potentially lead to significant improvements in teams' performance in CS:GO.

[23] (Brewer et al, 2022)

This study attempts to create a well calibrated CS:GO Win Prediction model that improves the accuracy of probability predictions.

A well calibrated win predictor is one where results that are predicted with an X% probability do indeed occur X% of the time. However, many models built for sports prediction give preference to overall accuracy of predictions instead of the accuracy of the probabilities. The authors attempt to apply techniques used in weather prediction to create a CS:GO prediction model that is well calibrated.

Two methods of improving model calibration have been investigated: sigmoid and isotonic calibration. Isotonic calibration was found to be extremely useful for many of the models tested, resulting in substantial improvements to the Calibration. The Multilayer Perceptron model was found to produce the best overall results, when combined with isotonic calibration, generating a well calibrated model that was capable of making refined, accurate predictions.

CHAPTER 3:
RESEARCH DESIGN

3.1 Objectives of the study

The primary objective of the study is to develop an **efficient and effective model** that can provide users with **insights into the probable outcome of rounds** in the game of Counter Strike: Global Offensive. This objective can be broken down into 3 sub-goals:

1. Develop a Model that can accurately **predict the outcome** of rounds in CS:GO based on data pertaining to that round.

2. Develop a Model that can accurately **predict the Win/Loss probability** of rounds in CS:GO based on round data.

3. Gain **Insights** into the factors determining success in a CS:GO round based on Models' features and weights

3.2 Statement of the problem

The major problems that need to be solved in order to meet the objectives of this study are:

1. **<u>Generate Accurate Predictions</u>** – This is the basic goal of the study. The predictions must be greater than trivial (i.e. >50%).
2. **<u>Generate Probability Predictions</u>** – Most models only output the final Prediction (Win or Lose) when given relevant data. However, our goal is to generate a Probability alongside this binary Prediction – for example, 63% chance of Winning, 33% chance of losing the round.
3. **<u>User Friendliness</u>** – The ultimate goal of the project is to create a model than can **readily be applied** and used by interested parties – analysts, teams, broadcasters etc. This means that the final model must be deployable and straight-forward to run
4. **<u>Interpretability</u>** – In order to derive insights from the models, they must be interpretable. Interpretable models are **models whose decision making process is readily accessible to users** - for instance from a decision tree you can easily extract decision rules and from

Multiple Linear Regression models you can determine the importance of each variable. This is important if actionable insights are to be extracted from the models.

3.3 Scope of the study

- The Scope of the Study is restricted to the game of Counter Strike: Global Offensive
- The analysis is conducted on Tournament Matches (High Skill Level) conducted in 2019-20, but the models developed should be capable of accurately predicting the outcome of matches outside this timeframe.
- The scope of the data is restricted to Round Data – data pertaining to the individual rounds. Overall Match data and data regarding teams & players are not considered.
- Scope is also restricted to developing accurate model. Analysis of the reasons why certain variables have an impact of round outcome comes falls outside the scope of this study.

3.4 Data Overview and Operational Definition

- **Dataset** - CS:GO Round Winner Classification (AKA CS:GO Round Snapshots)
    - **Source**: Kaggle.com
        - Dataset has 10/10 Usability Ranking in Kaggle.
        - Published by Christian Lillelund, Kaggle Datasets Expert and Research Assistant at Dept. of Engineering, Aarhus University
    - **Original Publisher**: Data was released by Skybox.GG, a CS:GO analytics software provider, as a part of their CS:GO AI Challenge.

- **Variables**: There are a total of 97 Features

i. 96 Independent Variables & 1 Dependent Variable (Round Outcome)

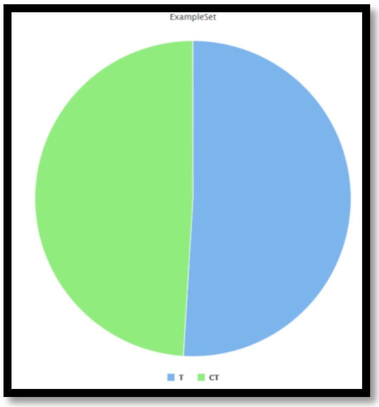
ii. 94 Numeric Variables (Integer) & 3 Nominal Variables

The 97 variables can be grouped into the following categories:

- **Weapons** (68 - 34 weapons x 2 teams) – these variables explain the different types and quantities of weapons in the possession of each team
- **Equipment** (17 – 8 items x 2 teams plus 1 special item) – these variables explain the what types of grenades, armor and other equipment the teams possess
- **Miscellaneous Round Data** (12) – These variables describe of state of the round and the teams at that particular time - such a Time Left, Number of Players alive, Health of Players etc.
- **Prediction Label** (1) – The Round Winner, which acts as the label for each observation. This is what the models aim to predict, using the 96 variables above.

- **Total Observations**: 122,410 snapshots of rounds
  - These observations are “statistical snapshots” taken approximately every 50 seconds during a round. The snapshots consist of various data points related to the round.
- **Pre-Processing** – The datasets has already been pre-processed and flattened by the publisher in order to improve readability and make it easier for algorithms to process
- **Label/Predicted Value** = Winner of the Round. Terrorists (T) or Counter-Terrorists (CT)

a. Distribution of Wins



The T:CT Win Ratio of the dataset is 51:49, indicating this is a balanced dataset

Operational Definitions:

- **T – Terrorist**, one of the two sides in a CS:GO match. Their objective is to either eliminate all players in the CT side, or plant and detonate a bomb before time runs out
- **CT – Counter Terrorist**, one of the two sides in a CS:GO match. Their objective is to either eliminate all players in the T side, or prevent the bomb from being detonated until time runs out.
- **Round**: In a round, two teams take turns playing on the two sides – CT and T – working to achieve their objectives before the 1:55 minute time limit runs out. Whichever team successfully completes their objective is termed the winner. There cannot be a draw in a round.
- **Match:** A set of up to 30 rounds, in which teams compete to reach the match point of 16 round wins. The team to first reach 16 round wins is deemed the winner of the match. Teams switch sides (CT & T) after the first 15 rounds.
- **In-Game Economy**: Each player has a personal balance of money which they can use to purchase weapons and equipment. The amount of money earned by a player each round is determined by their actions and the actions of their team. This Money and Purchasing system is referred to as the in-game economy.

3.5 Hypotheses

Not Applicable, as the objective of the study is create a predictive model, not test hypotheses,

3.6 Model Design

All modeling has been done on Rapidminer. RapidMiner is a data science software platform that provides an integrated environment for data preparation, machine learning, deep learning, text mining, and predictive analytics.

Rapidminer eliminates the need to manually program the modeling workflow. Instead, it allows users to create workflows and develop models through an intuitive drag-and-drop environment, akin to Tableau for Data Visualization.

Individual Model designs have been given in the next chapter, along with explanations.

3.7 Method of data collection

Secondary Data was used, as explained in the previous section. It was downloaded from Kaggle.com.

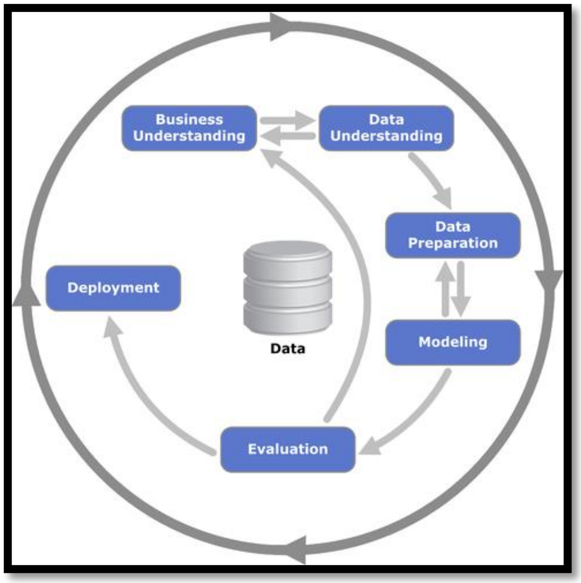
3.8 Sampling Type / size

All 122,410 observations were used in the training and testing process. The data was divided into a 70/30 Train-Test split using Stratified Sampling

3.9 Workflow Design

The CRISP-DM (Cross-industry standard process for data mining) has been used in this study CRISP-DM one of the most widely accepted data analytics frameworks. The framework describes 6 keys steps involved in the data analytics process –

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?



3.10 Limitations of the Study

- The study is restricted to the prediction of Rounds based solely on round data. Data on the players and teams has not been included. The addition of such data to the modeling process has a significant chance of improving the overall predictive power of the model.
- Only models that are not too computationally intensive were applied in this study, due to the limitation of my hardware. Therefore some potential useful models had to be excluded.

CHAPTER 4: ANALYSIS AND INTERPRETATION

4.1 Basis for Model Selection:

1. **Ability to be used for Classification** – As this is a classification problem, only models capable of classification were considered
2. **Ability to Process both Numeric and Nominal Data** – Necessary in order for all data to be included in model
3. **Computing Power Requirements** – Model training must not take an excessive amount of time and/or computing power

The following five Models satisfied the above requirements –

- 2x **Traditional Statistics Oriented Models**: Generalized Linear Model and Logistic Regression
- 2x **Machine Learning Models**: Decision Tree and Naïve Bayes
- 1x **Ensemble Model** – a model that combines 2 or more other models into a single predictive model

The following section will explain the models, their functioning and performance in predicting the outcome of rounds.

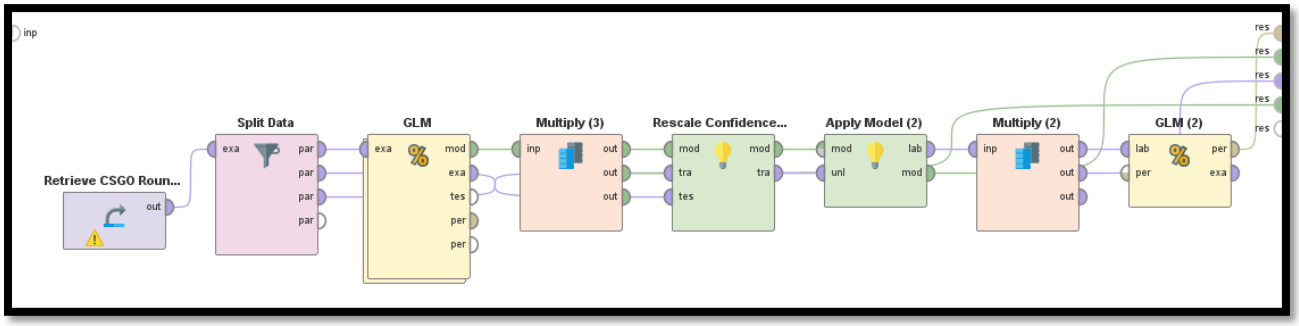
4.2 Performance of Models Tested

1. Generalized Linear Model

i. Definition

Generalized Linear Model (GLiM, or GLM) is an advanced statistical modeling technique formulated by John Nelder and Robert Wedderburn in 1972. It is an umbrella term that encompasses many other models, which allows the response variable y to have an error distribution other than a normal distribution. The models include Simple and Multiple Linear Regression, Logistic Regression, and Poisson Regression.

ii. Rapidminer Workflow

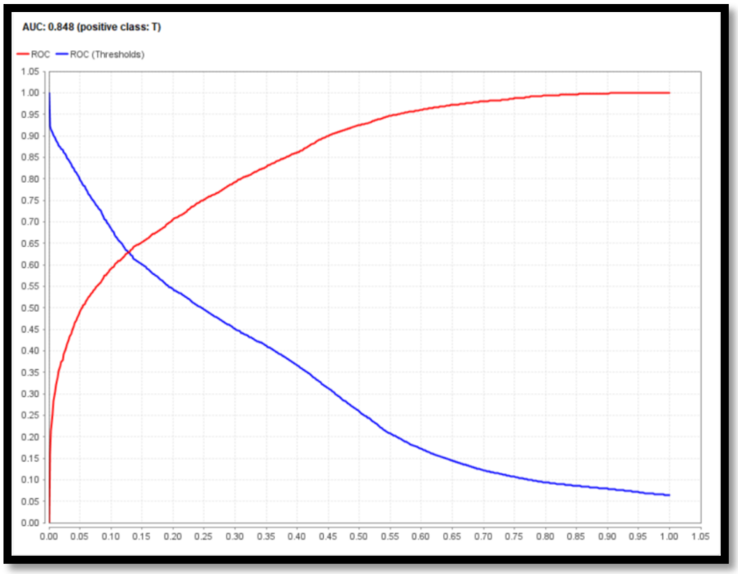


iii. **Accuracy**: 75.12%

a. Confusion Matrix

	true CT	true T	class precision
pred. CT	9054	3144	74.23%
pred. T	2947	9337	76.01%
class recall	75.44%	74.81%	

iv. **ROC Curve**: AUC = 0.848



v. Model-specific Statistics

a. Mean Square Error (Misclassification Rate) = 0.158

b. R-Squared value = 0.368

vi. Top 5 Features by Coefficient

a. All Features

Feature	Coefficient
ct_weapon_m249	3.250
t_weapon_bizon	1.097
t_weapon_p90	1.059
bomb_planted.True	1.025
t_weapon_mag7	0.850

b. Non-weapon Features

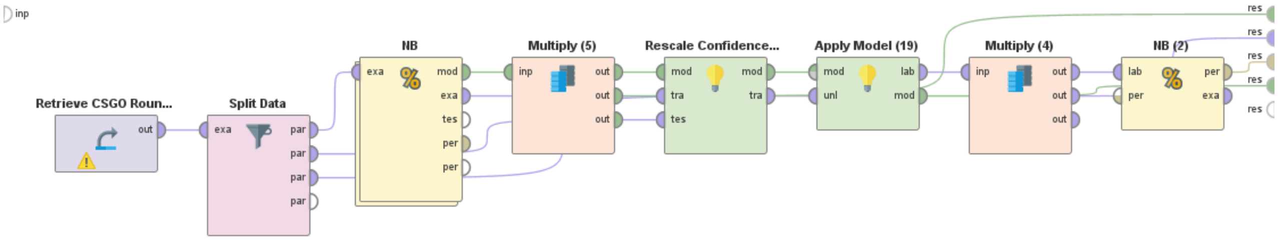
Feature	Coefficient
bomb_planted.True	1.025
t_players_alive	0.418
map.de_inferno	0.225
map.de_dust2	0.178
t_helmets	0.147

2. Naïve Bayes

i. Definition

Naive Bayes classifiers are a collection of classification algorithms based on **Bayes' Theorem**. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

ii. Rapidminer Workflow

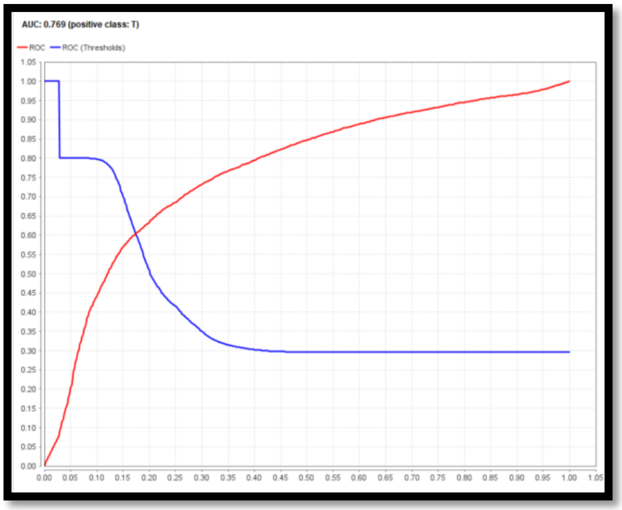


iii. **Accuracy**: 71.61%

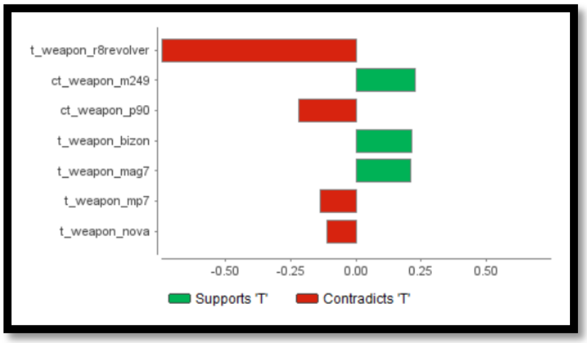
a. Confusion Matrix

	true CT	true T	class precision
pred. CT	9580	4529	67.90%
pred. T	2421	7952	76.66%
class recall	79.83%	63.71%	

iv. ROC Curve – AUC = 0.769



v. Top Features

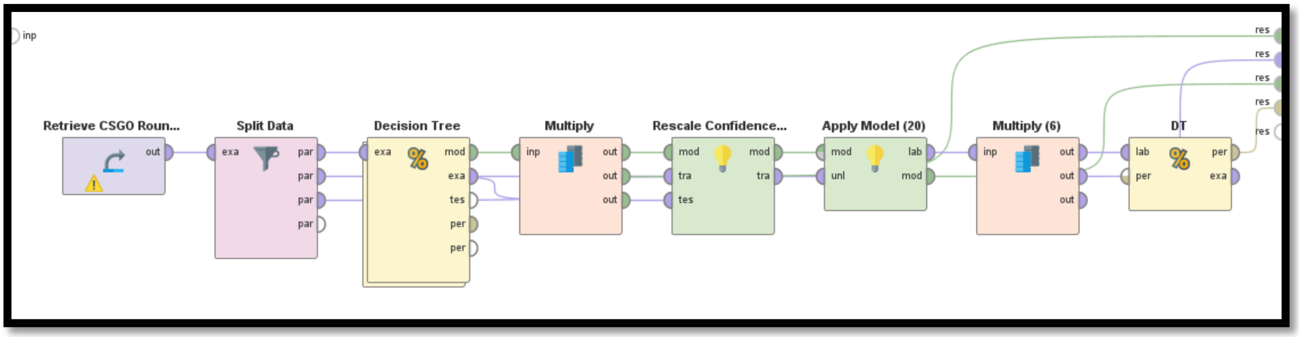


3. Decision Tree

i. Definition

A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The models arrive at a final prediction by testing and splitting until it reaches the optimal prediction.

ii. Rapidminer Workflow

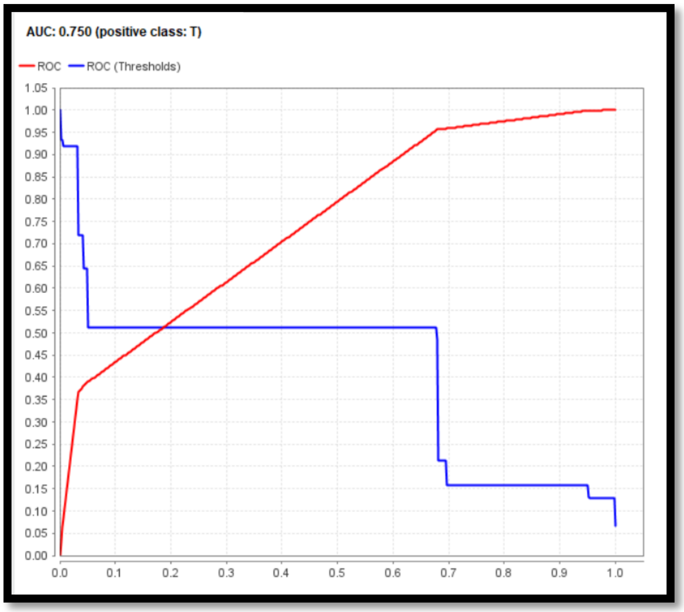


iii. Accuracy – 66.39%

a. Confusion Matrix

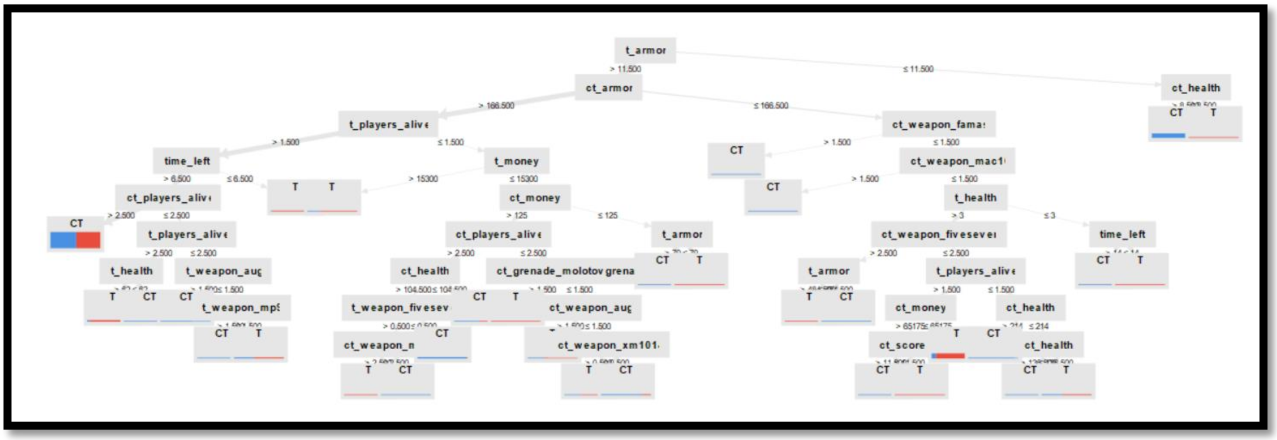
	true CT	true T	class precision
pred. CT	11396	7624	59.92%
pred. T	605	4857	88.92%
class recall	94.96%	38.92%	

iv. ROC Curve: AUC = 0.75



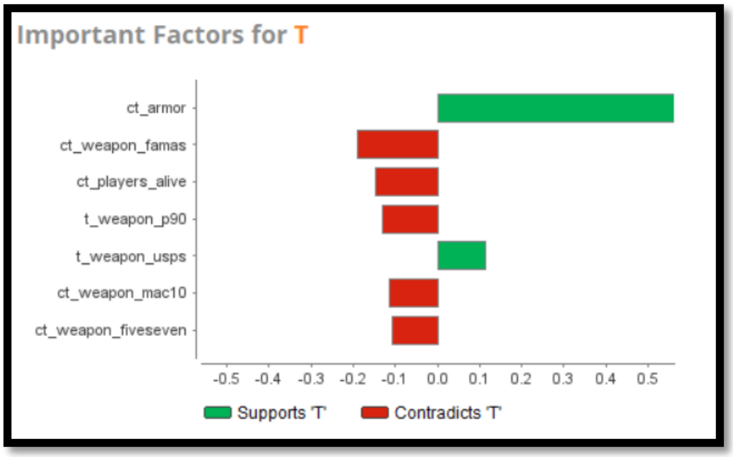
v. Model-specific Items

a.The Decision Tree

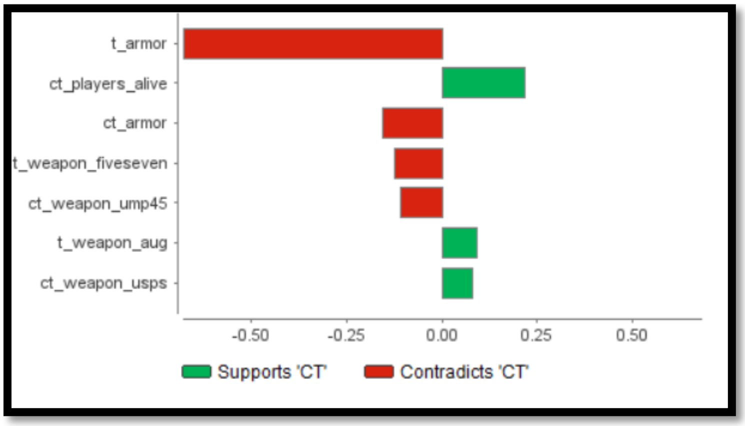


vi. Top Features

a. For T win



b. For CT Win

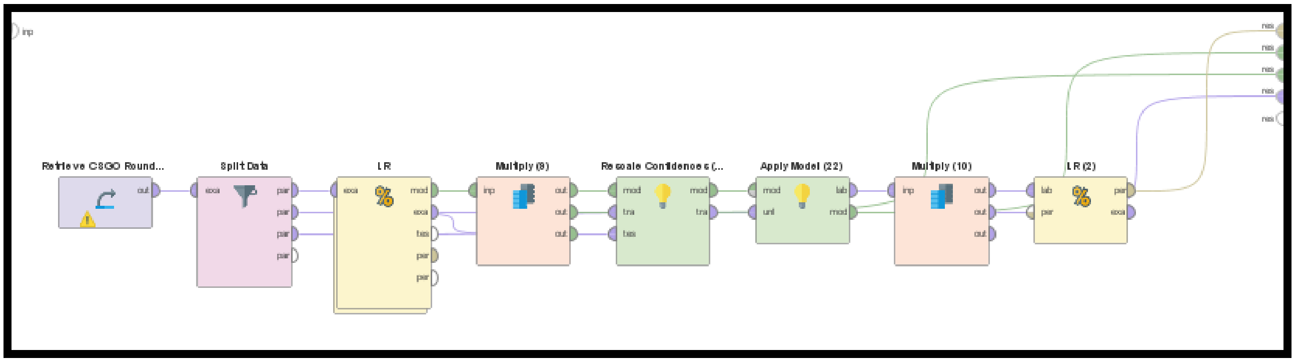


4. Logistic Regression

i. Definition

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category.

ii. Rapidminer Workflow

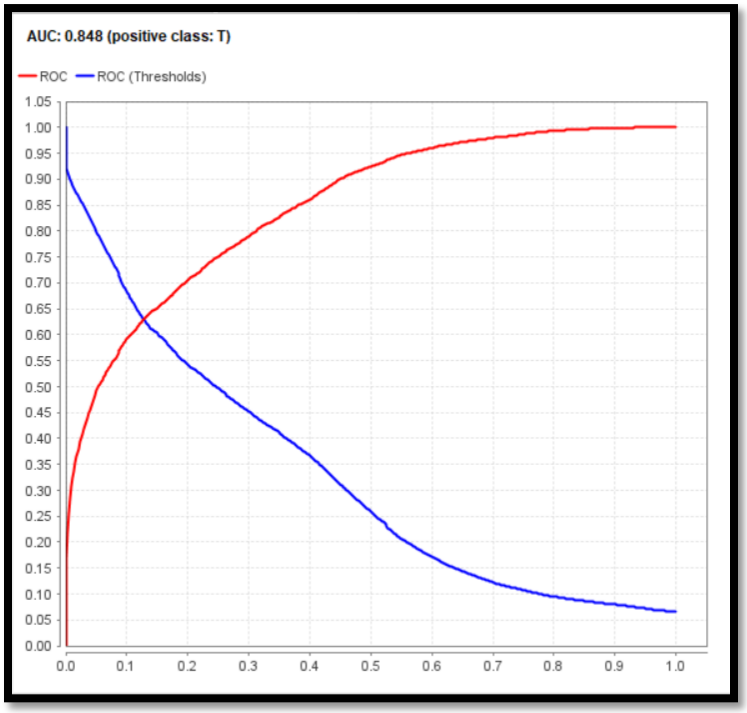


iii. **Accuracy**: 75.12%

a.<u>Confusion Matrix</u>

	true CT	true T	class precision
pred. CT	9053	3142	74.24%
pred. T	2948	9339	76.01%
class recall	75.44%	74.83%	

iv. **ROC Curve**: AUC = 0.848



v. Model-specific Statistics

a.Mean Square Error (Misclassification Rate) = 0.158

b.R-Squared value = 0.368

vi. Top 5 Features by Coefficient

a.All Features

Feature	**Coefficient**
ct_weapon_m249	6.513
t_weapon_mag7	2.110
t_weapon_bizon	1.279
t_weapon_p90	1.149
bomb_planted.True	1.081

b. Non-weapon Features

Feature	Coefficient
bomb_planted.True	1.081
t_players_alive	0.411
t_helmets	0.139
t_grenade_decoygrenade	0.117
ct_grenade_incendiarygrenade	0.094

5. Ensemble Model

i. Definition

Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data.

ii. Models used

Logistic Regression and Decision Tree models were stacked to create an Ensemble. LR has a high accuracy rate, whereas DT has High CT recall and T Precision. Both models have different strengths, making them suitable for combination.

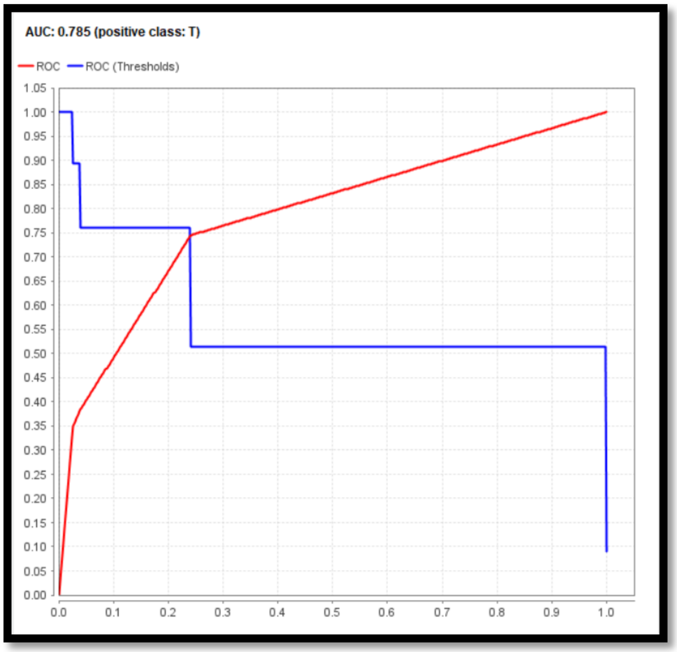
The final output model chosen was DT, due to its good accuracy rate and interpretability.

iii. Accuracy – 75.16%

a.Confusion Matrix

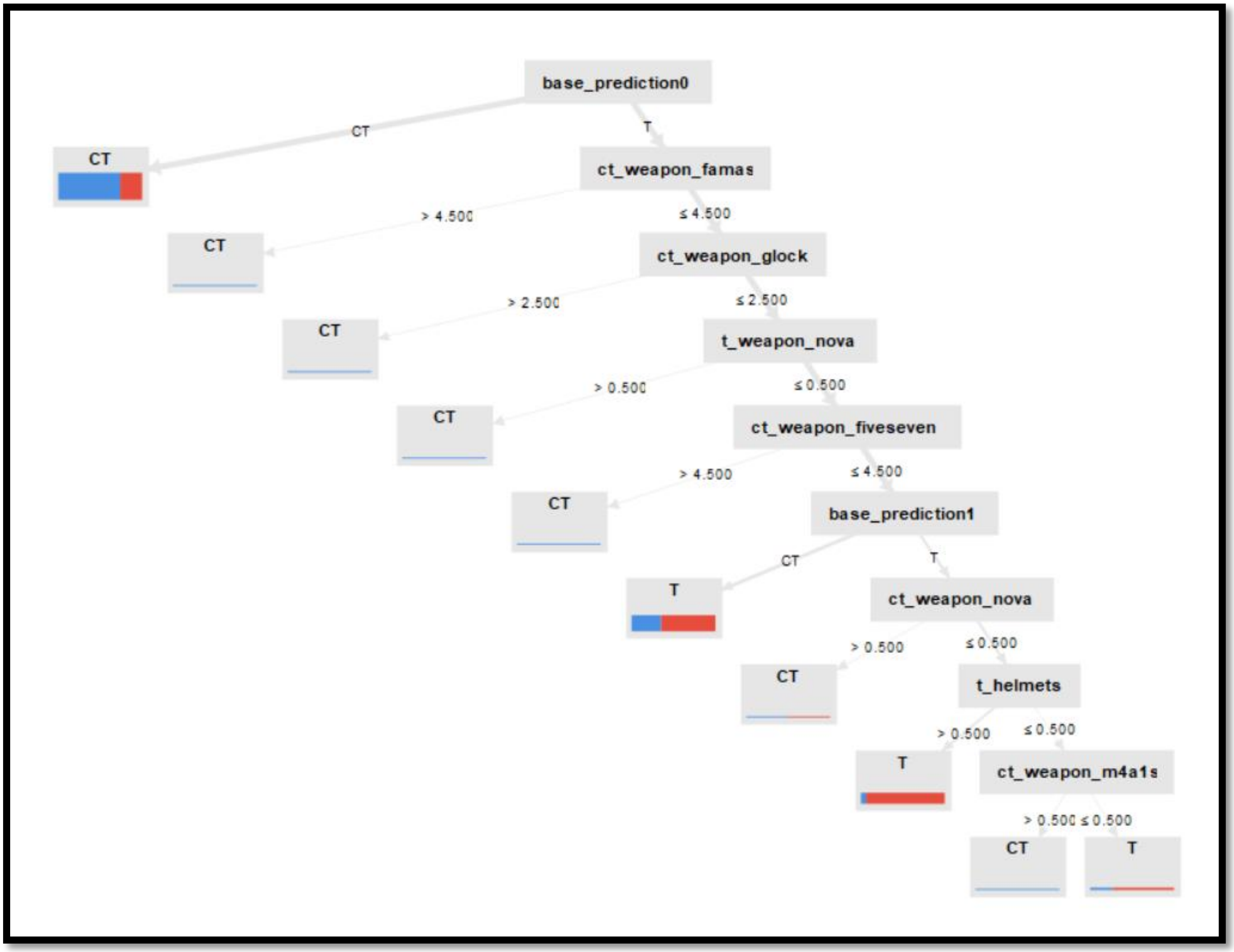
	true CT	true T	class precision
pred. CT	9104	3185	74.08%
pred. T	2897	9296	76.24%
class recall	75.86%	74.48%	

iv. ROC – AUC = 0.785



v. Model-Specific Items

a.Final Decision Tree



4.3 CALIBRATION PLOTS

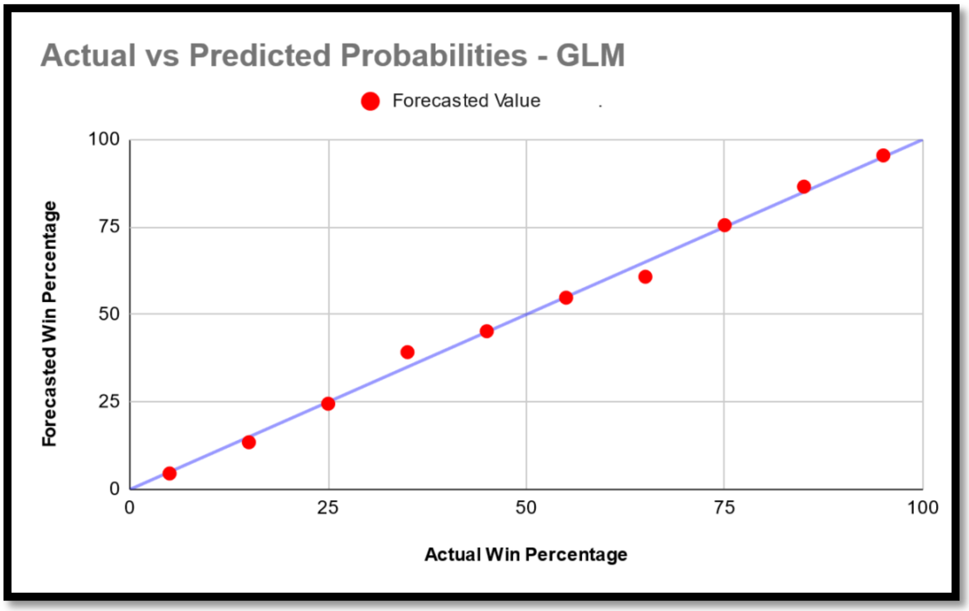
A calibration plot is a line-and-scatter plot which compares the observed probabilities of an event versus the predicted probabilities. A well calibrated predictor is one where results that are predicted with an X% probability do indeed occur X% of the time.

For example, take this study. The models output a particular prediction, CT or T, and a probability that said prediction will come to pass. – say 60% for CT or 70% for T. If we compare these predicted probabilities with the actual probabilities, we can test the calibration of the models.

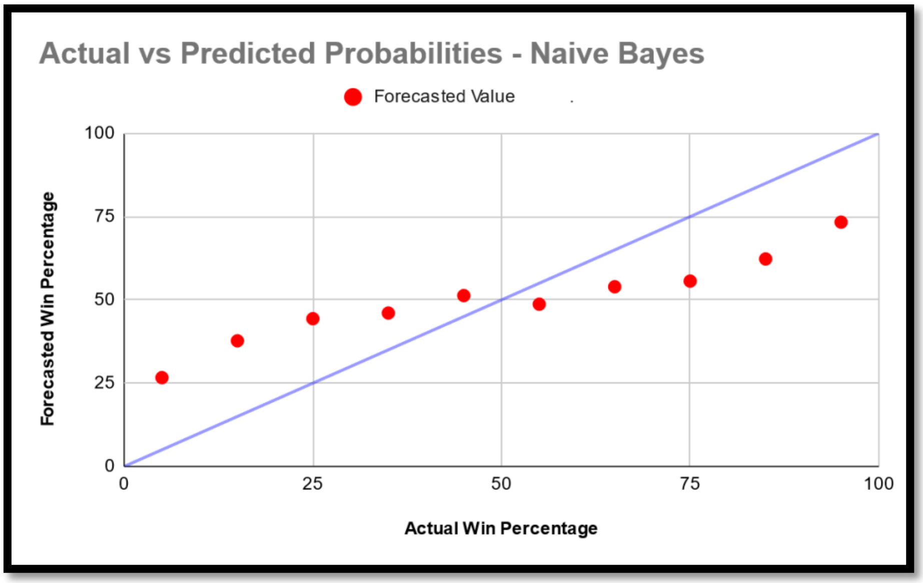
If say a model predicts that in situation the CT team will win 65% of the time, and the actual percentage is 62%, then we can say that the model is well calibrated. If this deviation is more, then it is said to have lower calibration.

CALIBRATION PLOTS of MODELS TESTED

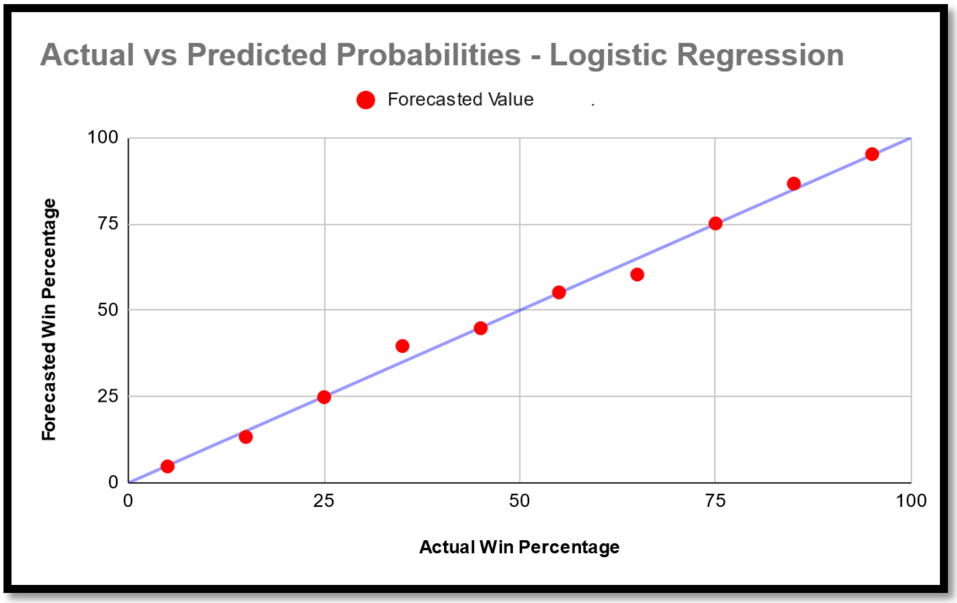
1. **GLM**



2. Naïve Bayes



3. Logistic Regression



4. Decision Tree & Ensemble

Not applicable as the output of both models are based on the Decision Tree algorithm, and the probabilities derived from DT were of no use.

CHAPTER 5:
SUMMARY OF FINDINGS

SUMMARY OF MODELS

		Models				
		GLM	Naive Bayes	Decision Tree	Logistic Regression	Ensemble Model (LR + DT)
Metrics	**Accuracy (%)**	75.12	71.61	66.39	75.12	**75.16**
	T Recall (%)	74.81	63.71	38.92	**74.83**	74.48
	CT Recall (%)	75.44	79.83	**94.96**	75.44	75.86
	T Precision (%)	76.01	76.66	**88.92**	76.01	76.24
	CT Precision (%)	74.23	67.9	59.92	**74.24**	74.08
	AUC	**0.848**	0.769	0.75	**0.848**	0.785
	Probability Calibration (Absolute Mean Deviation**) (%)**	**1.383**	16.21	-	1.385	-

Based on these finding we can conclude that –

i. The Ensemble Model is the most accurate predictor of CS: GO rounds. Decision Tree model is an expert in CT Recall and T Precision, but does not perform well in other metrics. By combining this flawed model with more balanced Logistic Regression, we are able to attain a high accuracy rate.

a. If the objective is to predict the ultimate round winner, the ensemble model is suggested

ii. In terms of the accuracy of the probability predictions, both GLM and Logistic regression perform well, with predicted probabilities only deviating from the actual probabilities by 1.38% on average. This makes both models suitable for Live Probability calculations, such as the ones seen in CS: GO Tournaments.

a. If the prediction of accurate probabilities is the objective, GLM and Logistic Regression Models are recommended

CHAPTER 6: RECOMMENDATIONS AND CONCLUSION

6.1 RECOMMENDATIONS

- **Time-honored Machine Learning models can be effective** as predicting the outcome of CS: GO rounds. Preference should be given to accuracy of probabilities and not accuracy of predictions, as throughout the majority of the round it is difficult to reliable predict the ultimate winner of said round. Similar to how weather predictions inform users of the probability of rain falling, a good round prediction model should inform users of the probability of a team's success in a round.
- The model should be kept as **simple and efficient** to run as possible. Models such as these are meant to be used in real-time. Therefore the more efficient a model is, the easier it is to do exactly that.
- Models such as the ones tested provide valuable information about the factors determining success in a round of CS:GO. Analysts, Broadcasters and Teams should **incorporate these insights** into their training and strategizing activities.

6.2 SCOPE FOR IMPROVEMENT

- In March of this year (2022), a paper on methods **to improve the calibration** of CS: GO predictive models was published. The recommendations made by the paper could potentially improve the performance of the models developed in this study.
- **Additional Features** with significant predictive power can be added onto or generated with the dataset used in this study to improve the accuracy of the models

6.3 SCOPE FOR EXPANSION

The proposed models, if errors are fixed and the models are refined, can be operationalized. For example, the models could be used to generate round win predictions for live games, and displayed during broadcasts of the same.

6.4 CONCLUSIONS

It is easy to build a working predictive model, but difficult to build a truly insightful and informative one. The models presented in this paper are simple, but they provide a reasonable level of accuracy and can be used to derive insights into the game of Counter Strike: Global Offensive. That was the goal of this study, and it has been sufficiently satisfied in the eyes of this study.

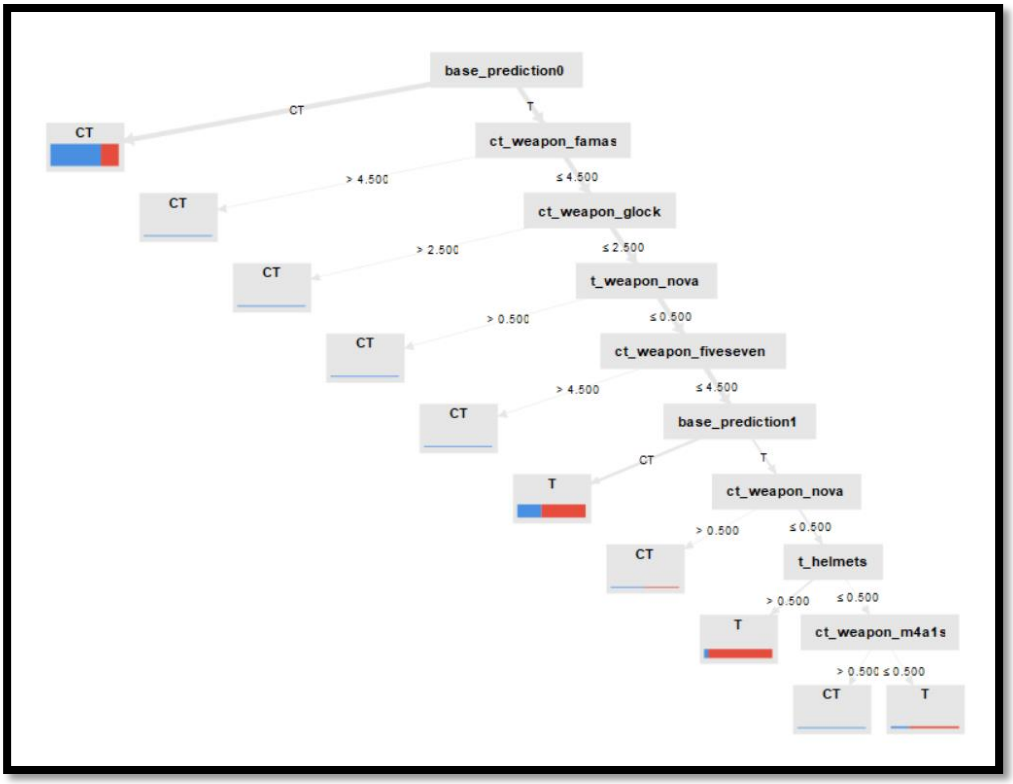
There is much room for improvement and for real-world implementation, which I hope to follow through on in the future.

APPENDIX

A. Summary of Models

		Models				
		GLM	**Naive Bayes**	**Decision Tree**	**Logistic Regression**	**Ensemble Model (LR + DT)**
Metrics	**Accuracy (%)**	75.12	71.61	66.39	75.12	**75.16**
	T Recall (%)	74.81	63.71	38.92	**74.83**	74.48
	CT Recall (%)	75.44	79.83	**94.96**	75.44	75.86
	T Precision (%)	76.01	76.66	**88.92**	76.01	76.24
	CT Precision (%)	74.23	67.9	59.92	**74.24**	74.08
	AUC	**0.848**	0.769	0.75	**0.848**	0.785
	Probability Calibration (Absolute Mean Deviation**) (%)**	**1.383**	16.21	-	1.385	-

B. Ensemble Model diagram



REFERENCES

[1] Steele, J. M. (2000). Eclipse Prediction in Mesopotamia. Archive for History of Exact Sciences, 54(5), 421–454. http://www.jstor.org/stable/41134091

[2] Borresen, J., & Lambert, M. I. (2009). The quantification of training load, the training response and the effect on performance. *Sports Medicine*, *39*(9), 779 - 795. 10.2165/11317780-000000000-00000

[3] Claudino, J. G., Capanema, D. d. O., de Souza, T. V., Serrão, J. C., Pereira, A. C. M., & Nassis, G. P. (2019). Current Approaches to the Use of Artificial Intelligence for Injury Risk Assessment and Performance Prediction in Team Sports: a Systematic Review. *Sports Medicine - Open*, *5*(1), 1-12. http://dx.doi.org/10.1186/s40798-019-0202-3

[4] Patel, S. (2021b, May 19). Esports vs Sports | Can Esports Compete with Traditional Sports? Retrieved from https://onlinegrad.syracuse.edu/blog/esports-to-with-traditional-sports/

[5] Hamari, J., & Sjöblom, M. (2017). What is eSports and why do people watch it? Internet Res., 27, 211-232.

[6] Statista. (2021, June 1). Worldwide eSports viewer numbers 2019–2024, by type. Retrieved from https://www.statista.com/statistics/490480/global-esports-audience-size-viewer-type/

[7] Wagner, M.G. (2006). On the Scientific Relevance of eSports. International Conference on Internet Computing.

[8] Reitman, J. G., Anderson-Coto, M. J., Wu, M., Lee, J. S., & Steinkuehler, C. (2020). Esports Research: A Literature Review. Games and Culture, 15(1), 32–50. https://doi.org/10.1177/1555412019840892

[9] Raquel Y.S. Aoki, Renato M. Assuncao, and Pedro O.S. Vaz de Melo. 2017. Luck is Hard to Beat: The Difficulty of Sports Prediction. In *Proceedings of the 23rd ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining (*KDD '17*). Association for Computing Machinery, New York, NY, USA, 1367–1376. DOI:https://doi.org/10.1145/3097983.3098045

[10] Makarov, I., Savostyanov, D.V., Litvyakov, B., & Ignatov, D. (2017). Predicting Winning Team and Probabilistic Ratings in "Dota 2" and "Counter-Strike: Global Offensive" Video Games. AIST.

[11] Hodge, V.J., Devlin, S., Sephton, N., Block, F., Cowling, P.I., & Drachen, A. (2021). Win Prediction in Multiplayer Esports: Live Professional Match Prediction. IEEE Transactions on Games, 13, 368-379.

[12] Wang, T. (2018). *Predictive Analysis on eSports Games: A Case Study on League of Legends (LoL) eSports Tournaments.* https://doi.org/10.17615/ez9n-t517

[13] Björklund, Arvid et al. (2018). "Predicting the outcome of CS:GO games using machine learning." Chalmers University of Technology.

[14] Švec, Ondřej (2022). "Predicting Counter-Strike Game Outcomes with Machine Learning.". Czech Technical University in Prague.

[15] Smerdov, A., Burnaev, E., & Somov, A. (2020). AI-enabled Prediction of eSports Player Performance Using the Data from Heterogeneous Sensors. ArXiv, abs/2012.03491.

[16] Xenopoulos, P., Doraiswamy, H., & Silva, C.T. (2020). Valuing Player Actions in Counter-Strike: Global Offensive. 2020 IEEE International Conference on Big Data (Big Data), 1283-1292.

[17] Petri, G., Stanley, M.H., Hon, A.B., Dong, A., Xenopoulos, P., & Silva, C. (2021). Bandit Modeling of Map Selection in Counter-Strike: Global Offensive. ArXiv, abs/2106.08888.

[18] McCabe, A., & Trevathan, J. (2008). Artificial Intelligence in Sports Prediction. Fifth International Conference on Information Technology: New Generations (itng 2008), 1194-1197.

[19] Min, B.K., Kim, J., Choe, C.J., Eom, H., & Ian, R. (2007). A Compound Framework for Sports Prediction: The Case Study of Football.

[20] Melentev, Nikita & Somov, Andrey & Burnaev, Evgeny & Strelnikova, Irina & Strelnikova, Galina & Melenteva, Elizaveta & Menshchikov, Alexander. (2020). eSports Players Professional Level and Tiredness Prediction using EEG and Machine Learning. 1-4. 10.1109/SENSORS47125.2020.9278704.

[21] Ghazali, N. F., Sanat, N., & As' ari, M. A. (2021). Esports Analytics on PlayerUnknown's Battlegrounds Player Placement Prediction using Machine Learning. International Journal of Human and Technology Interaction (IJHaTI), 5(1), 17-28.

[22] Yang, Z., Pan, Z., Wang, Y., Cai, D., Shi, S., Huang, S. L., ... & Liu, X. (2022). Interpretable Real-Time Win Prediction for Honor of Kings--a Popular Mobile MOBA Esport. IEEE Transactions on Games.

[23] Brewer, G., Demediuk, S., Drachen, A., Block, F., & Jackson, T. (2022). Creating Well Calibrated and Refined Win Prediction Models. *Available at SSRN 4054211*.

[24] Rubin, Allen, (2022), Predicting Round and Game Winners in CSGO. OSF Preprints, Center for Open Science.