GROUP B ASSIGNMENT

CS:GO Round Winner Classification

REPORT

This report summarizes the work, findings, and conclusions of Group B.

Albin S., Daniel B., Jorge R., Nicolas G., Tomas F., & Ckalib N

Machine Learning II

Table of Contents

Introduction

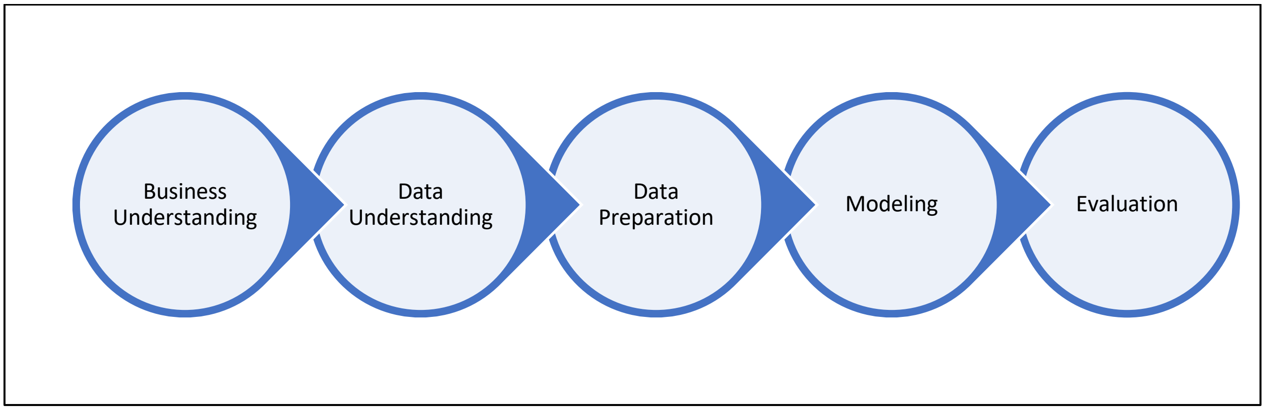
For the Group Assignment Project in Machine Learning II, Group B used their technical knowledge on Machine Learning to predict who wins in the tactical shooter game called, “Counter Strike: Global Offensive (GO).” This report, built with the Cross Industry Standard Process for Data Minding (CRISP-DM) in mind, serves as a summary of all our findings, insights, and conclusions from our analysis.

About the game

Counter Strike is multiplayer first-person shooter game initially released in 1999 on Windows. Due to its international success, CS:GO is already the fourth game released in 2012[1]. The game offers various maps on which two different teams play against each other, the Terrorists, and the Counter Terrorists. Both teams have the goal of eliminating the other team while simultaneously completing other objectives. The game offers different modes with different goals and difficulties, including challenges for players to manage their respective in-game economy to maximize their chances of success. Players can select and or buy different kits including various weapons, divided in a primary and secondary weapon, armor and bomb defusal kits[2].

Structure

As indicated in the beginning, the report and notebook follow the CRISP-DM methodology for data mining projects[3]. CRISP-DM is an open standard widely used, accepted, and originally developed by a consortium of over 200 organizations in 1996[4]. Although CRISP-DM is a six-phase methodology, our project will not require the final phase of deployment given the nature of the project. Thus, the authors followed the respective steps ranging from Business Understanding to Evaluation.



[1] https://blog.counter-strike.net/index.php/about/

[2] https://counterstrike.fandom.com/wiki/Category:Global_Offensive_game_modes

[3] https://www.forbes.com/sites/metabrown/2015/07/29/what-it-needs-to-know-about-the-data-mining-process/#2065f3a3515f

[4] The Knowledge Discovery Process, taught by Antonio Pita Lozano as part of the GMBD 2020 – 1

Business Understanding

During the Business Understanding Phase, Group B thoroughly identified, analyzed, and understood the business problem to translate it into an analytical problem. Group B heeded the warning of the late mathematician John W. Tukey before tackling the remainder of the project: “An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.” Accordingly, requirements and goals of the project were identified and listed below:

Requirements	**Goals**
If uncomfortable with Python (Pandas, NumPy, Matplotlib and SKLearn), use R	Predict who of the two teams wins in any given CS:GO game
Use Dataiku for Data Cleaning and Feature Engineering (if necessary)	Summarize work, findings, and conclusions in a written report
Build a binary classification model (Naïve Bayes, KNN, etc.)	
State relevant evaluation metrics	

Data Understanding

During the Data Understanding Phase, Group B collected the data to understand and explore it to extract information and insights. After collecting the data from Kaggle, an **exploratory data analysis (EDA)** was conducted for a thorough understanding of all of the features. EDA is the process of analyzing data to summarize their main characteristics through descriptive statistics and visualizations. It is during this process that the team can determine which libraries will be needed to further explore, prepare, and model the data.

As part of the EDA the variables distribution and correlation are explored by producing according histograms and heatmaps. Based on these observations, future steps relevant during the process of feature selection and engineering are added to the list of issues to be dealt with before building the model. To visualize the mentioned histograms, the most prevalent libraries seaborn and matplotlib are used.

Exploratory Data Analysis

After completing the EDA, Group B made the following observations about the dataset:

Shape of the Dataset	• 85,687 rows and 103 columns • Each row accounts for a snapshot during a round
Column types	• Float64 is the most common data type
Dataset Imperfections	• There are no missing values • The dataset shows no major imperfections
Correlations	• Most features show only little correlation. Those that do correlate will be explored and dealt with at a later stage
Distribution	• In the two most played maps, de_inferno and de_dust2, the terrorists won more games than the counter terrorists • If the bomb is planted (which does happen much less often than it is not), the terrorists group wins more often than counter terrorists • T win much more rounds when CT don´t buy any defuse kit • There are several weapons that are never used by the T or CT

Features	• Most features are related to the weapons used in the game, this means that a way to reduce dimensionality could be to group the weapons
Target	• The dependent variable, "round winner" is binary - either the terrorists or counter-terrorists won

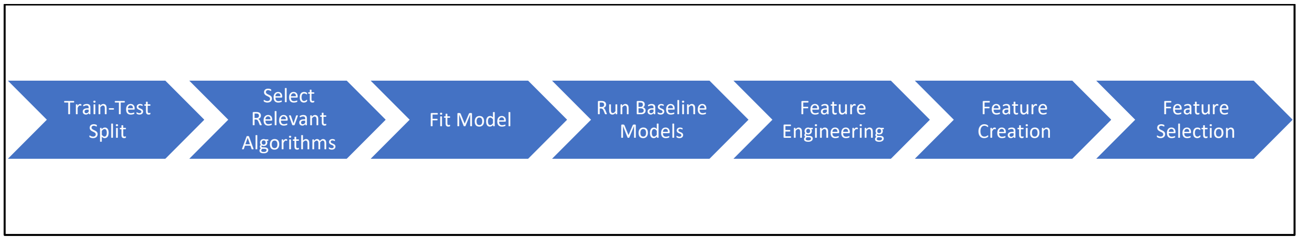
Data Preparation

During the Data Preparation Phase, Group B analyzed the data to decide if any features needed to be removed, created, or amended. Although the dataset we are working with is quite healthy, there are a few steps needed to have the data prepared for modeling:

Impute values	• Since none of the variables have any missing values, none had to be imputed
Outlier Check	• A total of 74 columns have shown some outliers. However, after checking those values, it was decided to keep all of them as they were deemed as plausible values in any Counter Strike game
Identify low frequency values & outliers	• 22 variables with low frequency (< 3) were identified • All those variables were weapons which were never, or almost never used. Accordingly, their removal will be considered in the feature selection stage
Scale numerical variables	• Using the StandardScaler from the sklearn library, all numerical variables were scaled to commensurate all numeric variables
Reduce dimensionality via PCA	• Using the PCA library from sklearn, a principal component analysis was applied to facilitate the classification • The resulting 4 principal components (PC) did not add any additional value, facilitating the classification. Thus, it was decided not to use the PCs in the final model
Clustering Attempt	• The team investigated whether or not certain combinations of guns yielded a favorable result via a clustering analysis; unfortunately, given the poor evaluation metrics of a K-Means analysis, we could not conclude that this would be beneficial to our analysis of CO:GO games

Modeling

During the Modeling Phase, Group B built analytical models to extract knowledge from the data and solve the business problem. Given that there are several models we have at our disposal to solve the business problem, Group B completed the following machine learning process for insightful modeling:



Train-Test Split

Select Relevant Algorithms

Fit Model

Run Baseline Models

Feature Engineering

Feature Creation

Feature Selection

After building a couple baseline models to have a rough estimate about the expected accuracy, we made the following decisions:

- Because the map seems to be an important feature, we tried to improve the encoding of this categorical variable. After testing Onehot, Target, and Catboost, One Hot Encoding was selected as the best encoding technique
- We created several weapon categories for each of the teams: sniper, heavy, assault, machinegun, shotgun, gun and grenades based on the information available
- The new categories and other existing features were used to create additional features representing the difference between the teams. For example, there are features for the difference in health, armor, money, helmets, number of players alive, and number of weapons in each category. This way, for example, we can evaluate if having +1 in awp weapons is a tactical advantage for any of the teams
- Finally, using Genetic Programming we tried to build new features automatically as a combination of the original and new built features

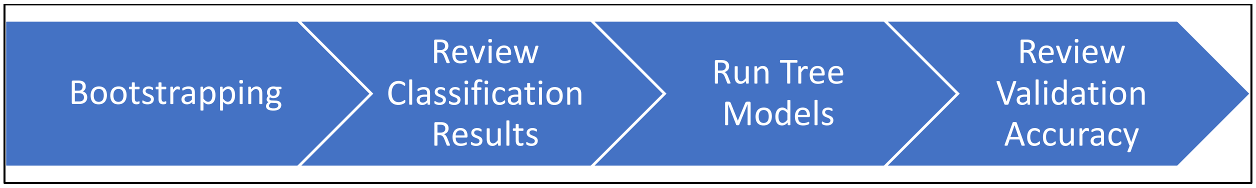
Once we had all the features created, we proceed to select the best features:

- Filter and remove all the weapons never used
- Analyze correlation of features but decide to keep them because we were running feature selection techniques afterwards
- Use a Random Forest Classifier to analyze feature importance. An important conclusion is that GP generated features are the most important but the model with them is less accurate. So we decided not to include any of these auto-generated features in the model
- Comparing feature selection based on RF feature importance and Recursive Feature Elimination, this second technique is performing better. So, after testing the right number of features to be selected we decide to use 60 features

So, after these steps we reduced the number of features from 97 in the original dataset to 60 and yielded roughly the same accuracy results.

Evaluation

During the Evaluation Phase, Group B evaluated the results to determine if the results addressed the business problem. Although our preliminary accuracies were quite good, we want to trying making them better by exploring how to reduce variance. Here is a list of tasks we completed for a full evaluation the model:



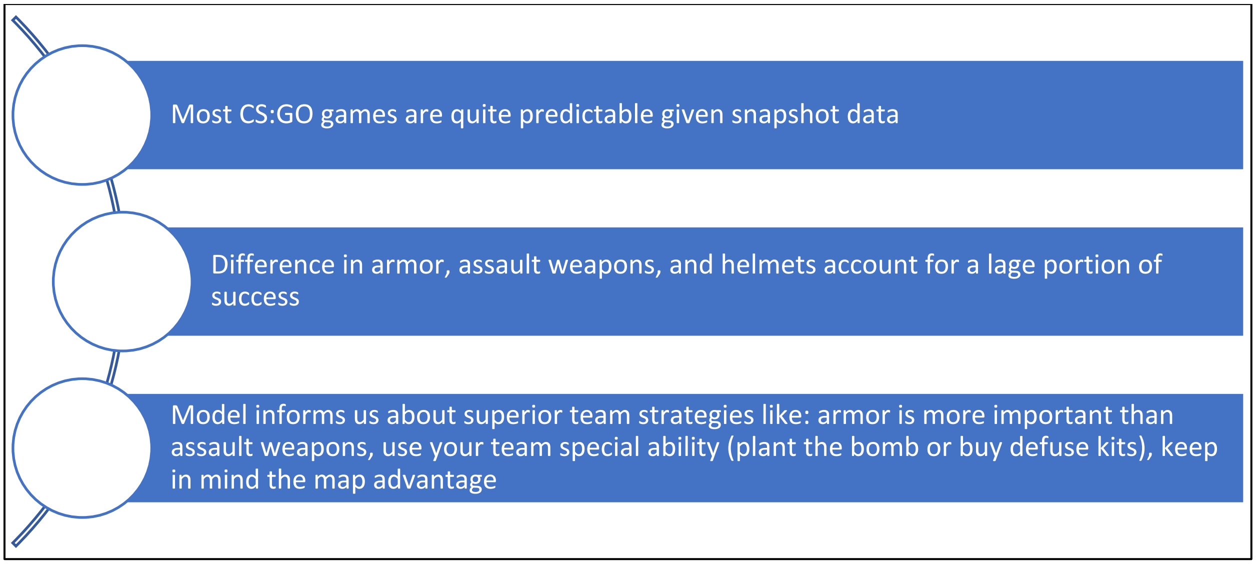
We validated our model using two different techniques:

- Using several samples from our dataset we want to make sure that our model is robust and don´t have much variance when we train and predict different sets
- We check that our predictions are good looking at the distribution of the GINI impurities

Finally, we tried to improve our model using two ensemble methods, basically these methods combined several models in order to improve the accuracy. Fortunately, with more time and effort after modeling, Group B achieved a Validation Accuracy score of roughly 85%, which refers to how close a measure is to the accepted value!

Conclusion

After following the CRISP-DM to complete this project, Group B was able to come to the following conclusions given the strength of our model:



- Most CS:GO games are quite predictable given snapshot data
- Difference in armor, assault weapons, and helmets account for a lage portion of success
- Model informs us about superior team strategies like: armor is more important than assault weapons, use your team special ability (plant the bomb or buy defuse kits), keep in mind the map advantage