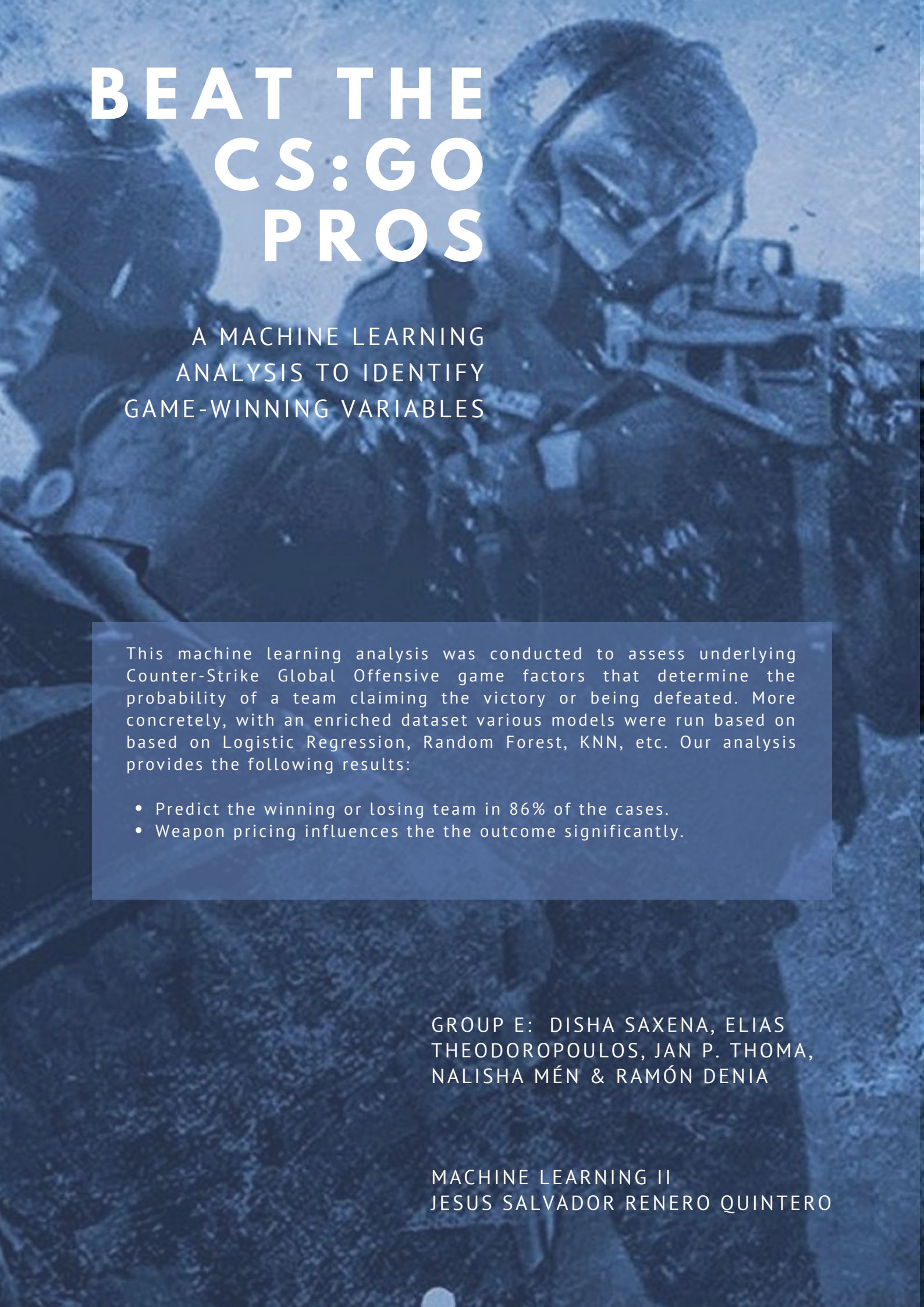


BEAT THE CS:GO PROS

A MACHINE LEARNING ANALYSIS TO IDENTIFY GAME-WINNING VARIABLES

This machine learning analysis was conducted to assess underlying Counter-Strike Global Offensive game factors that determine the probability of a team claiming the victory or being defeated. More concretely, with an enriched dataset various models were run based on based on Logistic Regression, Random Forest, KNN, etc. Our analysis provides the following results:

- Predict the winning or losing team in 86% of the cases.
- Weapon pricing influences the the outcome significantly.

GROUP E: DISHA SAXENA, ELIAS THEODOROPOULOS, JAN P. THOMA, NALISHA MÉN & RAMÓN DENIA

MACHINE LEARNING II
JESUS SALVADOR RENERO QUINTERO

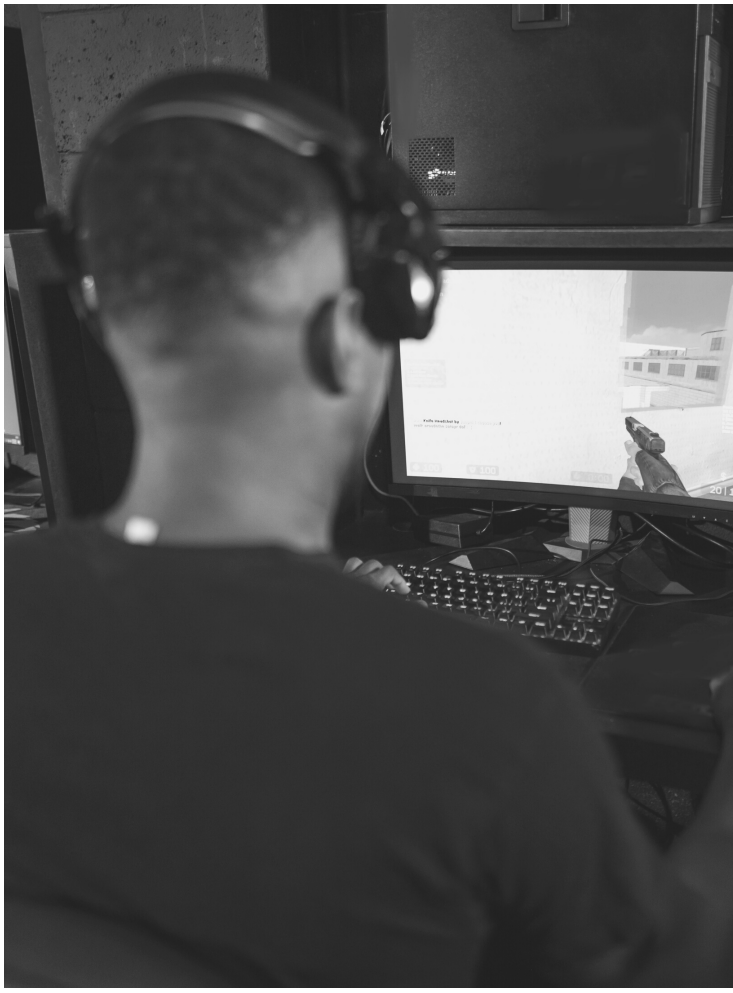
A GAME OF PURE SKILL?

IDENTIFYING GAME-WINNING VARIABLES WITH MACHINE LEARNING TECHNIQUES

With an estimated 11 million players per month since its release in 2012, Counter-Strike Global Offensive (CS:GO) is one of the most popular tactical shooter games. The gameplay is straightforward: two teams, terrorists and counter-terrorists, with five players on each, have 1min 55 seconds to win one round of a match. Within the given timeframe, terrorists (T) win by either eliminating the counter-terrorist team or ensuring the explosion of their bomb. The counter-terrorists (CT), on the other hand, win by eliminating all terrorists or diffusing their planted bomb.

While the gameplay is straightforward and tactical shooter games' results significantly depend on players' skills, it is unclear whether other factors contribute to a victory or defeat of a team. Based on our machine learning analysis of a Kaggle dataset, this report will unveil the most important factors unrelated to skill to increase your team's chances of winning a CS:GO match. The dataset consists of 97927 observations with 96 features from high-level tournament games in 2019 and 2020. To ensure a machine learning pipeline that works well in practice, we followed a two-step data preparation approach.

Our ML models identify weapon pricing as most significant.



First, we explored the features given in the original dataset, followed by replacing categorical values with numerical values. Moreover, we performed one-hot encoding on the feature "map". Second, with our deep understanding of the game, we decided that more valuable information can be extracted using feature engineering. As a result, we can rely on type categories and price categories for the large number of weapon systems.

With the enriched dataset, various models were run, including logistic regression models, random forest, support vector machines, and KNN. Through cross-validation and further tuning of the models, we ensure that our final model provides clear recommendations on game-winning features.

MODELLING WITH GAMING EXPERIENCE

With the enriched dataset allowed for the creation of three final datasets in which the individual weapon features are replaced to reduce the number of features while increasing the information provided by the dataset. As data cleaning was only necessary to a limited extent the following datasets were used for modelling.

The Price Category Dataset
In this dataset single weapons systems were replaced with weapon price categories. In total, there are four different price categories consisting of a low ($200-$1500), medium($1500-$2800), upper ($2800-$4100),and high-end ($4100-$5400) category. This categorization emphasizes the importance of game economics in professional matches in which victories are awarded with significant amounts of money. Before running the models the dataset which contained 32 features was split into a test and training set. Furthermore, feature scaling was performed on the non-boolean values.

The Type Category Dataset
In this dataset single weapons systems were replaced with weapon type categories. In total, there are five different type categories consisting of equipment, pistols, SMG, rifles and heavy weapons. This categorization emphasizes the importance of weapon characteristics in professional matches in which maps and playing style are impacted by the weapon of choice.. Before running the models the dataset which contained 32 features was split into a test and training set. Furthermore, feature scaling was performed on the non-boolean values.

The Weapon PCA Dataset
In this dataset single weapons systems were replaced with pca-based features. The PCA Dataset provides us with a statistical approach to reduce the number of weapon features without losing significant information. Before running the models the dataset which contained 32 features was split into a test and training set. Furthermore, feature scaling was performed on the non-boolean values.

3
Different Datasets Engineered

20+
New Variables Engineered

15
Models Run in Total

THE PRICE CATEGORY DATASET

```
The accuracy for LG is: 0.7473665791776029
[[10447  3442]
 [ 3777 10909]] 0.7473665791776029 0.7601560866838548
              precision    recall  f1-score   support

           0       0.73      0.75      0.74     13889
           1       0.76      0.74      0.75     14686

    accuracy                           0.75     28575
   macro avg       0.75      0.75      0.75     28575
weighted avg       0.75      0.75      0.75     28575
```

The baseline model was run as a Logistic Regression. Although it provided satisfactory results with an accuracy of 75% further models were run with the objective to increase all metrics, including precision, recall and f1-score. Hence, KNN, Decision Trees, Random Forest and Support Vector Machines were run. The best results are shown below.

With KNN another model was examined. This algorithm the increased the accuracy significantly by 10 percentage points to a level of 85%. Similarly, the other metrics show significant improvement with the examination of the confusion matrix.

```
The accuracy for KNN is: 0.8517235345581802
[[11803  2086]
 [ 2151 12535]] 0.8517235345581802 0.8573285001025922
              precision    recall  f1-score   support

           0       0.85      0.85      0.85     13889
           1       0.86      0.85      0.86     14686

    accuracy                           0.85     28575
   macro avg       0.85      0.85      0.85     28575
weighted avg       0.85      0.85      0.85     28575
```

```
The accuracy for RFC is: 0.8619072615923009
[[12064  1825]
 [ 2121 12565]] 0.8619072615923009 0.8731758165392633
              precision    recall  f1-score   support

           0       0.85      0.87      0.86     13889
           1       0.87      0.86      0.86     14686

    accuracy                           0.86     28575
   macro avg       0.86      0.86      0.86     28575
weighted avg       0.86      0.86      0.86     28575
```

The Random Forest Model provided the highest accuracy of all models run. Moreover, it surpassed the KNN model for all other metrics, including precision and recall. Hence, the final model chosen is based on Random Forest and the Price Category Dataset. Cross-validation and tuning of the models were run afterwards.

THE ECONOMICS OF THE GAME

FINAL CONLUSION

Applying cross-validation and further tuning the models their performance could be further increased. On the right-hand side the correlation matrix for KNN with Manhattan Metric showcases this step. The high performance of the model based on the Price Category Dataset is a strong indicator for the importance of economics in a professional game of CS:GO. Money can almost exclusively be gained through victories which provides a team with better weapons and in turn giving them an advantage in the following round.