CS6220 Project Report

Delphinus

Benjamin Ruzzo, Danny Godwin, Divyanshu Sharma, James Cook, Suneeth Ravi

Abstract

We aim at employing data mining methods to analyze Dota 2 match statistics data and to get insights into how the player and match metrics correlate, and what metrics can be used to predict the probability of winning for a team. We compare the effectiveness of different features to predict the win to get the strongest predictors. We also plan to analyze match results and obtain a set of optimal hero picks that maximize the probability of winning. To provide a comprehensive summary of the game, we also plan to analyze and visualize the player demographics, professional matches, and other interesting and informative metrics. We use the dota-2-matches dataset which is a collection of different match metrics files and gain information about 50,000 ranked ladder matches, team composition, player levels, and match statistics. We observe that creating a prediction model with all post-match statistics gives a highly accurate win prediction estimate, so we analyze the effectiveness of different metrics and subsets of the dataset to predict match outcome in advance.

Introduction

Predicting match outcomes in sporting events has always been a prominent machine learning research area. Sports analytics has become a popular tool in many professional sports games to aid decision-making.

Dota 2 is a multiplayer online battle arena game which has had a massive impact in the gaming industry. With the growth of the game internationally, the massive fanbase it has achieved, and the “International Tournament 2021” prize pool reaching $40M, And the prize of professional Dota 2 tournaments has already passed $80 million by June 2021. It has sparked a lot of research interest around analyzing the game data to predict match results and provide optimal combinations to increase the probability of winning the game. It's likely that "online analytics" will be valuable for participants in professional online sports contests, live streaming media covering these competitions, and online sports game creators, just as sports analytics has been employed in professional sports decision-making.

Before we talk about how and what we are working on it is important to understand about the game and its features work. Dota 2 is a Valve-developed video game. The game includes ten players divided into two teams called 'Dire' and 'Radiant' (each team has five players shown in figure 1), and the goal is to defend your ancient tower from the enemy. Each match takes place on a map divided in half by a river.

Each player selects a hero from a pool of 115 options (shown in figure 2). In the game, there are not only heroes but also creeps who appear every half minute. Buildings can also throw fire at the closest enemy to the building during an attack; hence, players should avoid approaching the tower unless they have amazing power. We can see the full illustration of all heroes, creeps, buildings, and maps in a screenshot taken during the game in figure 3.

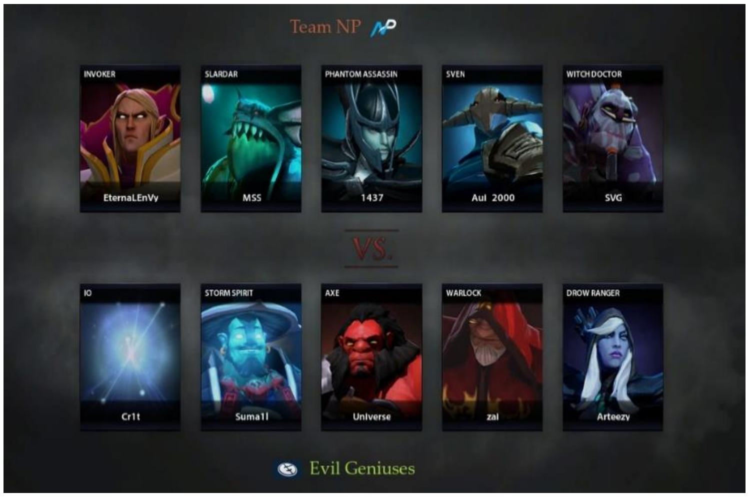


Figure 1

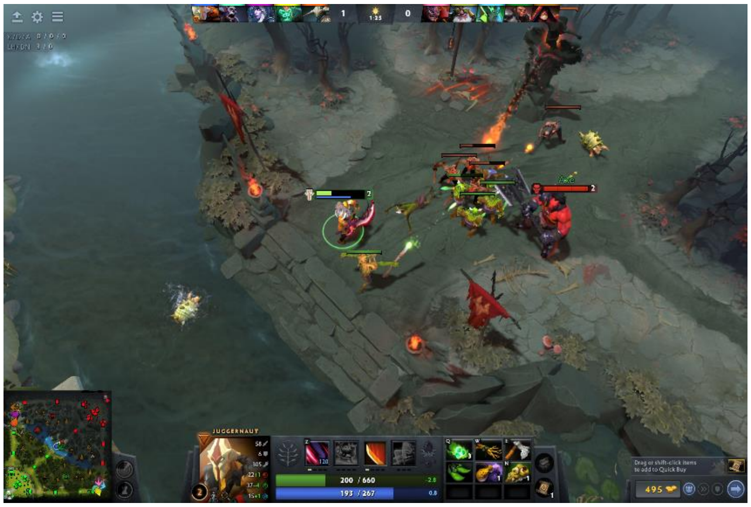
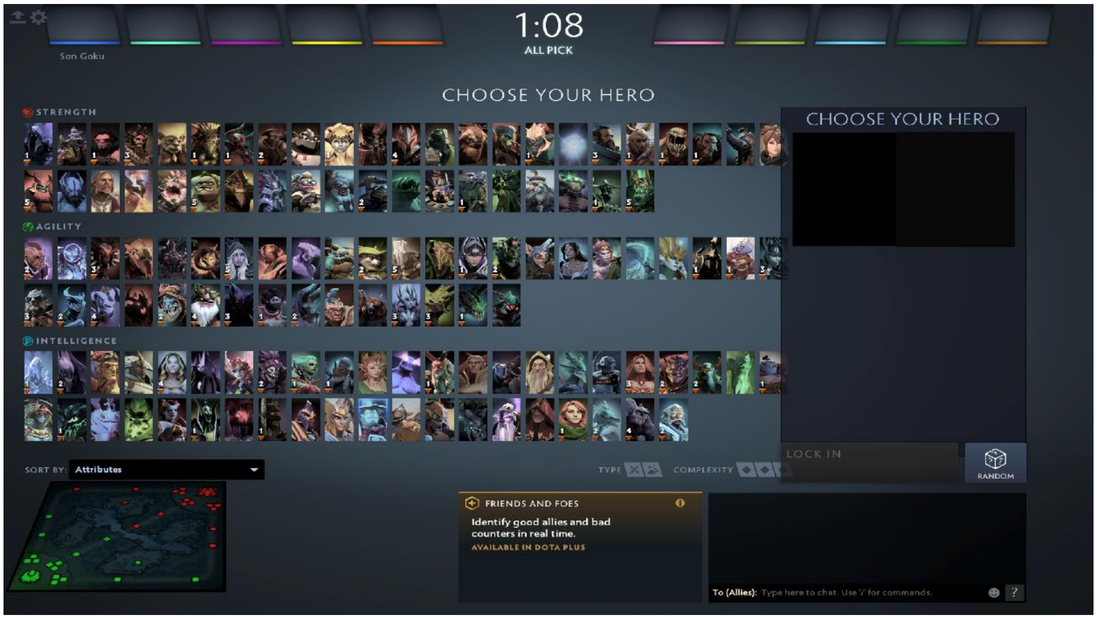


Figure 2



Identify good allies and bad counters in real time.

AVAILABLE IN DOTA PLUS

To (Allies): Type here to chat. Use '/' for commands.

Figure 3

In this paper we are focusing on visualizations of the exploratory analysis on the online matches data, professional matches data and demographics. In the online sports game Dota 2, we attempt to predict the winning side in a match between two teams named "Radiant" and "Dire," each of which has five players. Each player chooses a unique "hero" character to control in the contest before it begins. The hero can gain gold or gain experience to level up as the game goes by fighting other heroes. A team's winning criteria is to annihilate the opposing team's fountain. We intend to predict the game outcome by using different sets of features and data, including all details about the match, the post-match team stats, team composition, player skill level, and the economic strength of the team among other feature sets. We also look at player demographics, hero pick rates and correlation of different metrics in a match and provide visualizations and analysis of the data.

Related Work

There has been a lot of work done in sports in the past to predict the outcome of a game in different types of sports. But when it comes to online sports, we need to move away from the traditional approaches, for example using Bayesian networks to predict the outcome of football games and utilize the factors like zone change, distribution of team members, and time series clustering to focus on hero classifying hero positional role and hero identifier based on play style and performance. Previous work was also done on various games, for example League of Legends, by using clustered player behavior and learning the optimal team composition. Then it used team composition-based features to predict the outcome of the game.

Logistic regression and k-nearest neighbors have been used previously to recommend hero selections that would maximize the team’s winning probability against the opponent team. The features employed in earlier research, however, are merely hero choices and/or hero victory rates, which are drawn from a very restricted set of game data. Furthermore, the data from the real-time gameplay is completely ignored. As a result, expanding the feature set that will remedy these flaws is an important aspect of this research. We intend to include damage done to buildings during the game, the gold level of the team, the variation in the strength of a team's economy and player levels to contribute as dynamic factors contributing to the better prediction of the game. We approach the prediction problem as both post-match as well as real-time prediction and perform a comprehensive analysis of the features available in the dataset and how strongly they correlate with the win probability.

Datasets

- **Primary:** Dota 2 Matches - https://www.kaggle.com/datasets/devinanzelmo/dota-2-matches
    - This dataset contains 50000 ranked ladder matches from the Dota 2 data dump created by Opendota.
    - It comprises of 18 csv files containing match, player, outcomes, ratings and chat.
- **Secondary:** Dota 2 Professional Games Hero Picks - https://www.kaggle.com/datasets/soulreaper328/dota-2-professional-games-hero-picks
    - Dota 2 hero pick and match result for professional matches

Methodology

We started by looking at different collections of Dota 2 match metrics to help us analyze player behavior, strategies, and overall interesting metrics about way the game plays out. Our focus was on finding important predictor variables to help us predict a match outcome given the match details and player information. Along with the primary problem of predicting the winning team, we also wanted to look at other aspects of the game based on auxiliary datasets. We analyzed the difference in play styles across different regions in the world, the best team composition for wins and which heroes are the most popular among players

To load and aggregate the matches dataset, we use read in each csv files from the Dota matches dataset into a data frame using pandas. For each of these Dataframes, we aggregate and explore the data feature using the "describe" method to output descriptive statistics such as the number of records, mean and standard deviation of the dataset. We also used the appropriate plots to visualize relevant features of each dataset.

Professional hero picks

We used the professional matches hero picks dataset to analyze the hero choices teams make. Team composition is a particularly important part of Dota 2. Heroes in Dota can be categorized into distinct roles based on the abilities they have, when the abilities become effective, and base hero statistics. To maximize the probability of winning, an optimal team is composed of a combination of heroes that encompass all roles, which allows the players to support each other through the game. For instance, heroes with the "carry" role are less useful at the start of the match, but other "support" heroes are useful at lower levels. An effective team composition should have both types of heroes so that the support players can help level up during early game, and the carries can take over once their heroes are levelled up.

Hero pick rates also give us an insight into the perceived strength of a hero that players might have, or the popularity of the hero. We also analyze the win rates of different heroes in a professional setting and try to infer if there is an optimal team combination that drastically raises the chances of winning a match.

The dataset contains the id of each hero in the winning and losing teams for 71.4k matches. To analyze the popularity and win rates of each hero, we process the data to get the total number of matches each hero featured in, and the number of matches won. This helps us get the popularity of the hero, as well as the win rate. Based on this, we can analyze if there is a correlation between the popularity of a hero to its win rate. We further analyzed the team compositions and found out that owing to the considerable number of heroes available to be picked, we have a lot of unique team combinations with similar win rates. We can therefore infer that there is not a single combination of heroes that maximizes the probability of winning, and more depends on the skill of the player that picks the hero rather than the hero's abilities. Additionally, we can also infer that the heroes are well balanced, wherein picking a hero does not provide an obvious advantage to any team.

Game duration across regions

The duration of a match helps us infer the play style of the players of a region. Shorter match times may point to a more aggressive compared to a slower and more calculated play style. Different regions of the world differ in this metric in a lot of online games, and we aim to see if this applies to Dota 2 as well.

The match dataset includes the cluster or region id for where the server on which the match was played is located. We merged the two datasets and analyzed the distribution of match durations across regions. This metric also helps us identify the maturity of the player base, i.e., how good individuals are at the game overall. We perform a t-test to determine statistical significance of the difference in mean across regions.

Predicting match winner

For predicting the match outcome based on statistics that we have about the matches and players, we implemented different models on subsets of the data as well as subsets of features, to get both post-match and mid-match predictions.

- **All features:** We selected as many features from the matches and players datasets as we deemed useful. Each match has corresponding team information with xp, gold, kills, deaths etc. per player. We merged the two data together by calculating the per team summaries of the individual metrics. Based on the correlation plots, we saw that a lot of these features are pretty strong predictors for the team win, as the average number of kills, gold and xp for a team has a high chance of predicting the outcome of the match.
  This classifies as a post-match analysis of the data. We trained two linear classification models, a Logistic Regressor and a Support Vector Classifier, over a training dataset of 37.5k matches and 30 features per match. As expected, both the classifiers achieve near perfect accuracy and ROC AUC (Area Under the Curve) as they can correctly predict the outcome of matches based on player statistics.

- **Economy:** The economy of a team is an incredibly significant part of an RTS game. The amount of in-game currency a team has dictates which and how many abilities they can level up and what additional perks they can unlock, among other things. This currency

takes the form of gold in Dota 2. Gold is the currency used to buy items or instantly revive your hero. Gold can be earned from killing heroes, creeps, or buildings. Much of a player's in-game impact can be gauged through their net worth.
The team that has a better economy has an edge throughout the match. Tracking the amount of gold each team has during a match tells us a lot about the team's probability of winning**.** We aggregated the gold levels per team and added it to the matches data. Initial exploration showed that there was a strong correlation between the gold a team has and the team's probability of winning. We also analyzed the variation of the effectiveness of the predictor for varying time duration through the game.

- **Barracks kill:** Barracks are buildings that spawn Lane Creeps and are defended by Tier 3 towers. There are two barracks for each lane. The time of destruction of barracks can be a helpful marker to predict the outcome of the match, as whichever team gets the first barracks kill demonstrates better teamwork and gets an advantage as it gets easier control over lanes as well as limits the farming capabilities of the opposing heroes.
  We can get the timestamp of the first barracks kill by picking the first time the 'CHAT_MESSAGE_BARRACKS_KILL' message pops up in chat and merging the data with the match data to verify which team got the first kill. Based on our analysis, the first team to get a barracks kill increases their win probability by a big margin. The correlation weakens as time passes in game.

Leaving Players

Players leaving the game is a serious concern in team competitive games. In casual games, the missing players are usually filled, but in professional or ranked games the slot is just left empty, leaving the team down a member and at a serious disadvantage. Predicting if a player will leave is an important task and can work to maintain team composition and avoid the handicap. Our dataset includes a feature, 'leaver_status' which breaks down the different reasons for leaving:

0 - NONE - finished match, no abandon.

1 - DISCONNECTED - player DC, no abandon.

2 - DISCONNECTED_TOO_LONG - player DC > 5min, abandoned.

3 - ABANDONED - player DC, clicked leave, abandoned.

4 - AFK - player AFK, abandoned.

5 - NEVER_CONNECTED - player never connected, no abandon.

6 - NEVER_CONNECTED_TOO_LONG - player took too long to connect, no abandon.

A new binary feature 'leave' was created by setting the categories labeled 'abandoned' as 1 and the rest as 0. Using these two features we created predictive models based on the player's match statistics.

Dota 2 Professional Games Hero Picks

To understand how the professional players choose heroes based on various factors we took the Dota 2 Professional Matches data set which consists of a csv file with winning and losing hero

Id’s for 75k matches and a JSON file containing the hero Id along with their names. We first converted the JSON file into CSV so that we could integrate the datasets. Next we calculated the number of wins and losses for each hero and calculated the total matches played by a hero, their win and loss rates, and popularity. With this information we utilized Autoviz to generate meaningful comparisons and correlations between the metrics No of Wins, No of losses, Win Rate, Lose Rate, Total Matches and Popularity which help in determining the accurate factors of choosing a hero.

Dota 2 Professional Games Hero Picks:

We first divided the data set into winning heroes and lost heroes and calculated the no of wins for winning heroes dataframe and the no of losses for losing hero. Later we took the JSON file with hero id and hero name and converted it into a CSV file. Now we merged the winning hero, losing hero and the hero's name dataframes with respect to the hero id which is available in all these dataframes. Now as we have a hero wins and losses in a single dataframe we can calculate the total matches a specific hero has played and then determine the Win rate and lose rate. Popularity of a hero was calculated and AutoViz was used to plot various relations between these factors to determine the factors for picking a specific hero.

Player Skill Prediction

The skill of each player on the team will have a huge impact on which team wins the game. We wanted to see if we could predict the skill of a player based on the number of games they have won. The data set “Player_Ratings” had five features, the account id of the player, total wins, total matches, trueskill_mu, and trueskill_sigma. The trueskill_mu feature is the skill rating of the player, while the trueskill_sigma is the estimated variance of the skill rating. We then created a new feature called win rate, which is simply the number of wins divided by the number of games played. We also wanted to make the skill rating easier to understand, so we binned the ratings into levels of beginner, intermediate, and pro, and gave them the number representation of 0, 1, 2 respectively. This made it easier to categorize players based on skill.

Results

Professional hero picks

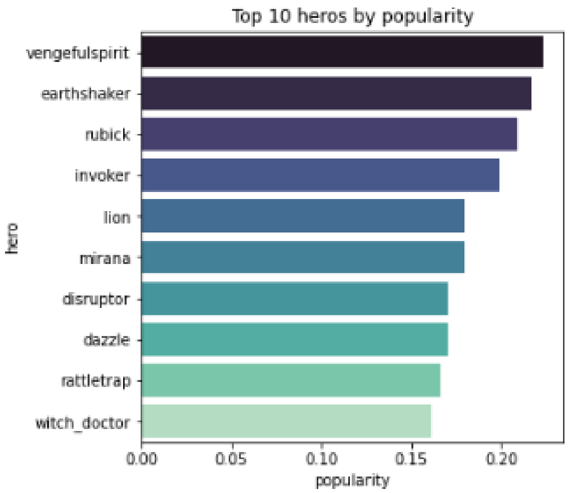


Figure 4: Top 10 heroes by popularity

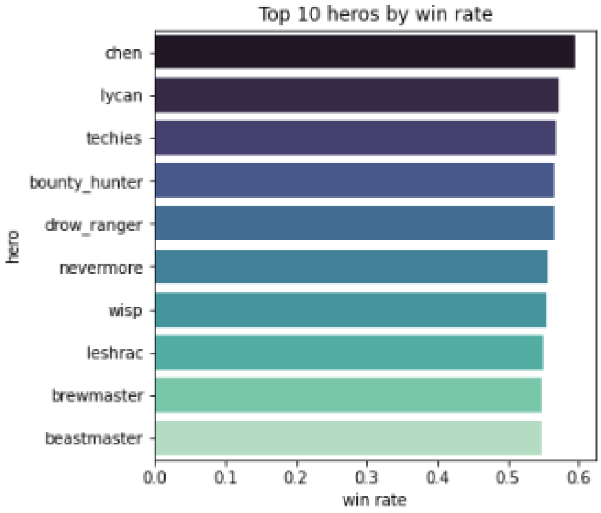


Figure 5: Top 10 heroes by win rate

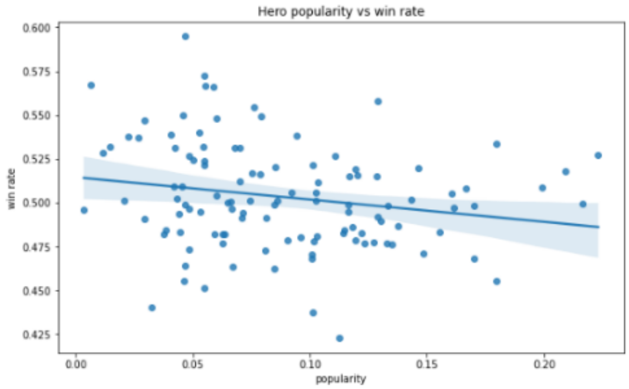
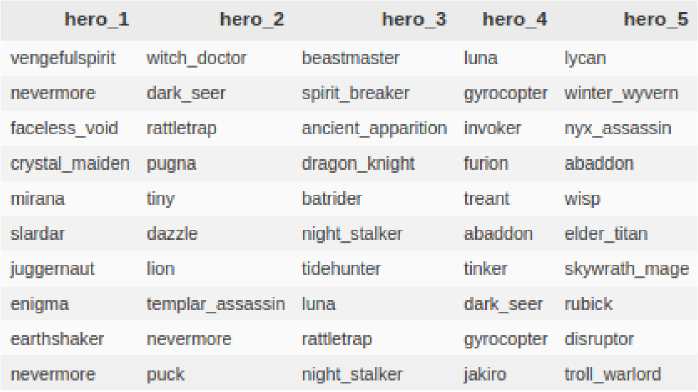


Figure 6: Correlation of popularity and win rate

Analyzing the heroes' picks dataset helps us get a list of the 10 most picked heroes (Figure 4), and top 10 heroes by win rate (Figure 5). We also analyze the correlation between the popularity of a hero and its pick rate (Figure 6) to see if a player's perception of hero strength translates to wins. We get a Pearson’s correlation coefficient of –0.2 with a p-value of 0.03.



hero_1	hero_2	hero_3	hero_4	hero_5
vengefulspirit	witch_doctor	beastmaster	luna	lycan
nevermore	dark_seer	spirit_breaker	gyrocopter	winter_wyvern
faceless_void	rattletrap	ancient_apparition	invoker	nyx_assassin
crystal_maiden	pugna	dragon_knight	furion	abaddon
mirana	tiny	batrider	treant	wisp
slardar	dazzle	night_stalker	abaddon	elder_titan
juggernaut	lion	tidehunter	tinker	skywrath_mage
enigma	templar_assassin	luna	dark_seer	rubick
earthshaker	nevermore	rattletrap	gyrocopter	disruptor
nevermore	puck	night_stalker	jakiro	troll_warlord

Figure 4: Top 10 team compositions by win rate

Additionally, we are also able to get the composition of the team that has the maximum win percentage.

Association Analysis

Using the hero data, we can also create association rules to see frequent pairings. Using apriori and FP Growth algorithms. Using the match heroes the top 5 support and top 5 confidence rules.

Hero(es)	Support
Shadow Fiend, Windranger	0.13124
Slardar, Windranger	0.09875
Earthshaker, Windranger	0.09009
Shadow Fiend, Slardar	0.08765
Tusk, Windranger	0.08745

Hero(es)	Confidence
{Spectre, Slardar} -> {Windranger}	0.44161
{Lich, Slardar} -> {Windranger}	0.44139
{Shadow Demon} -> {Windranger}	0.44027
{Winter Wyren, Anti-Mage} -> {Windranger}	0.43654
{Anti-Mage, Bounty Hunter} -> {Windranger}	0.43612

Separating the heroes into the two teams we get the following rules:

Hero(es)	Support

Dazzle, Windranger	0.01747
Rubick, Windranger	0.01706
Windranger, Winter Wyren	0.01598
Windranger, Wraith King	0.01596
Lion, Windranger	0.01451

Hero(es)	Confidence
{Omniknight} -> {Windranger}	0.20875
{Bounty Hunter} -> {Windranger}	0.20564
{Dazzle} -> {Windranger}	0.20451
{Rubick} -> {Windranger}	0.20387
{Winter Wyren} -> {Windranger}	0.20182

Game duration across regions

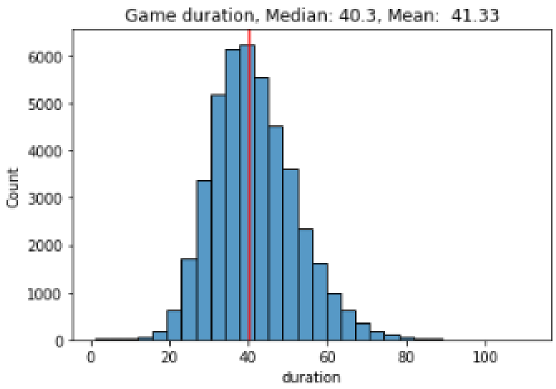


Figure 5: Game duration distribution

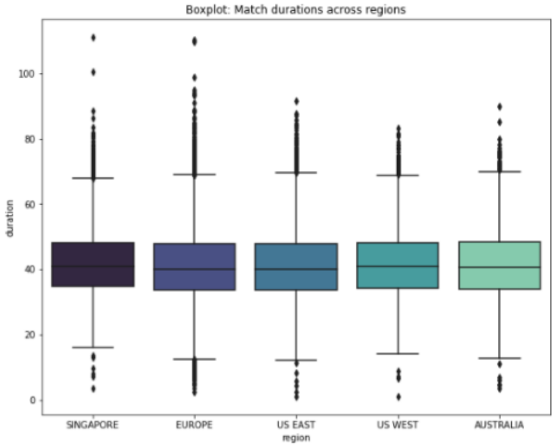


Figure 6: Boxplot of region-wise duration

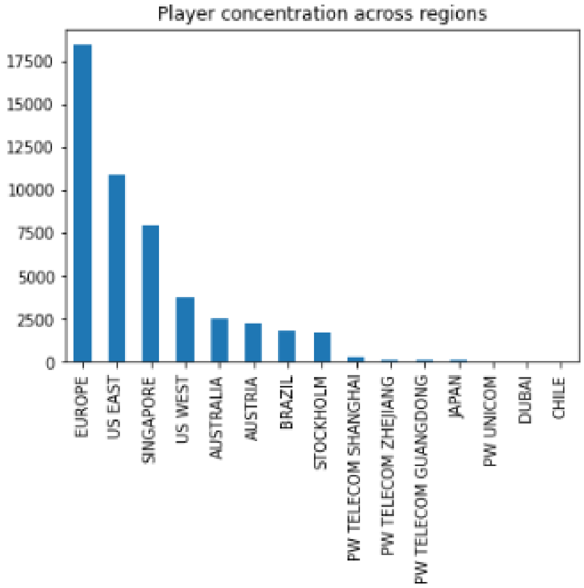


Figure 7: Player concentration across regions

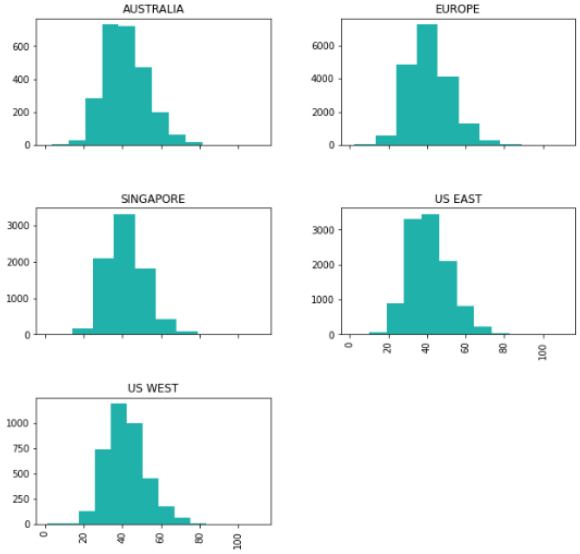
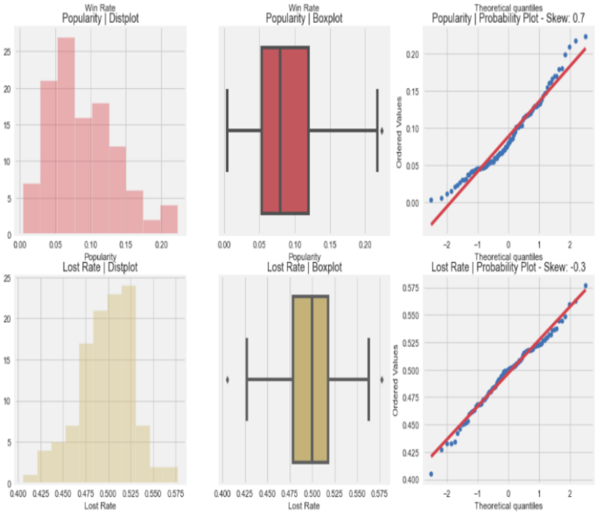


Figure 8: Distribution of durations for selected regions

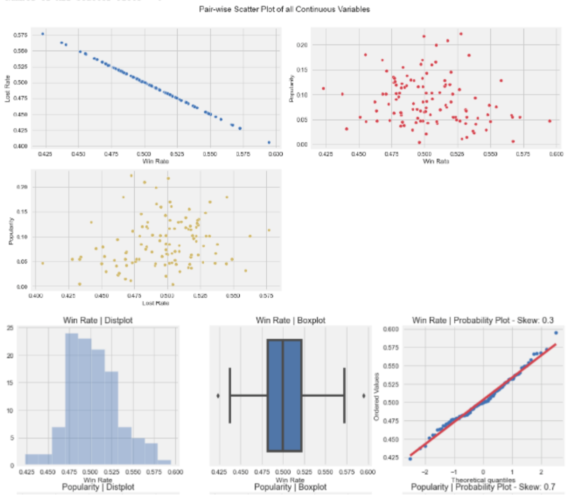
Region	**Mean Duration**
US East	41.1919
US West	41.7739
Europe	41.0831
Australia	41.5428
Singapore	41.8151

Using the matches data, we obtained the mean game duration for matches across different regions. For analysis, we shortlisted the top 5 most populated regions to see if there was a significant difference between match durations. We performed a pairwise t-test with the null hypothesis that the distributions are the same, to get the statistical significance of the variation in distribution means and to see which regions were significantly different in terms of duration.

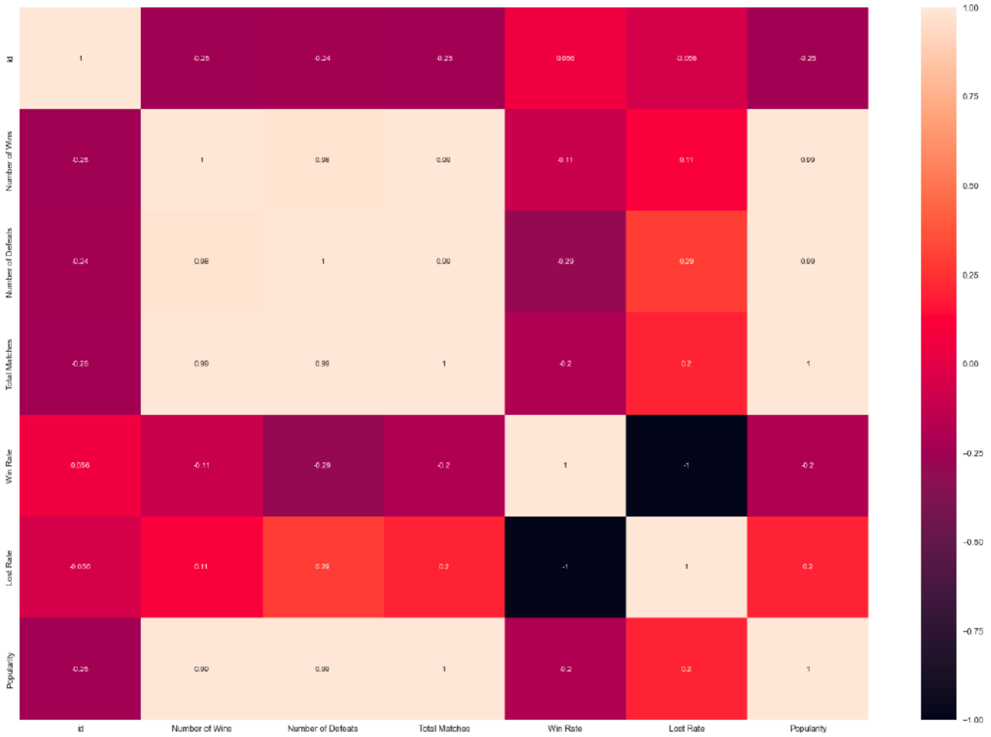
We also used AutoViz on the data generated from the code “***Dota 2 Professional Games Hero Picks”*** and also found some interesting metrics and correlations from it.



```
Number of All Scatter Plots = 6
```



Heatmap of all Continuous Variables including target =



Predicting match winner

- **All features**

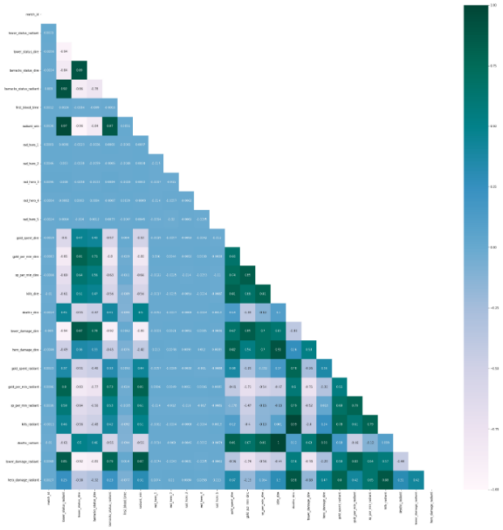


Figure 9: Heatmap of correlation between all features

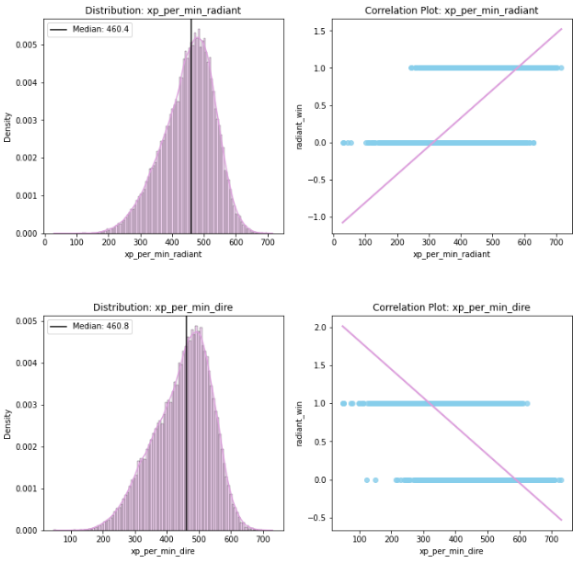


Figure 10: Correlation between opponent team features

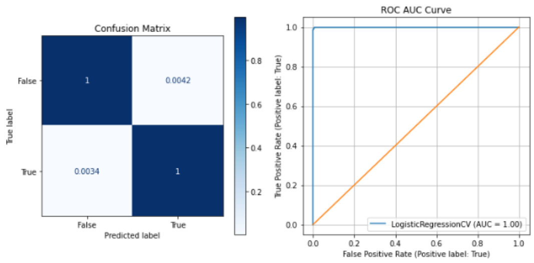


Figure 11: Confusion matric and ROC AUC of linear classifier

Using a combination of matches and players dataset, we were able to get features that had strong correlation with the win probability, as illustrated in figure, where the ‘xp_per_min’ is positively correlated for the winning team and negatively correlated for the losing team. A linear classifier can correctly model these relationships, evident from the near perfect ROC AUC.

- **Economy**

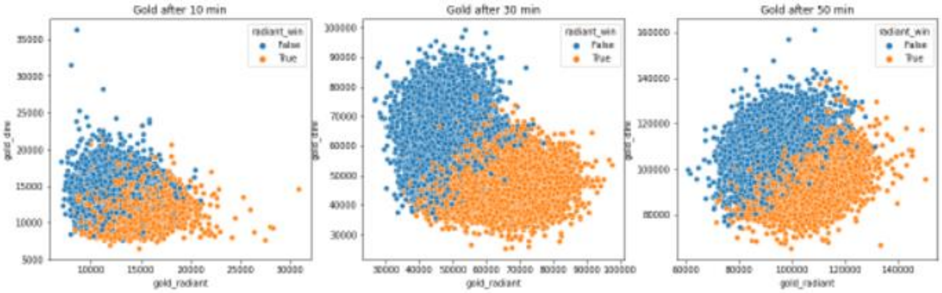


Figure 12: Scatterplot of team gold levels for selected duration of matches

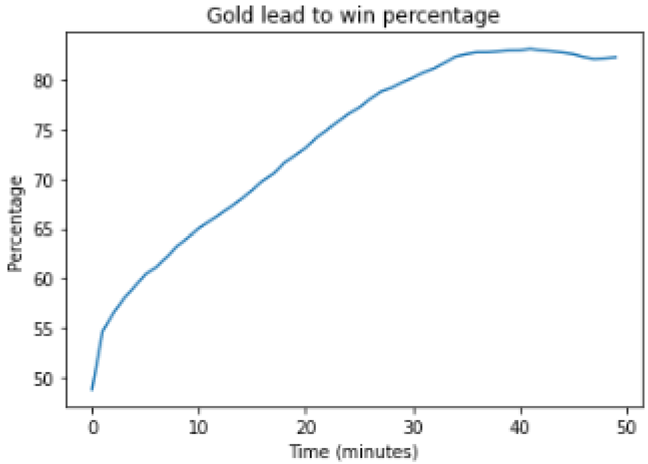


Figure 13: Percentage of teams in the gold lead that win the match

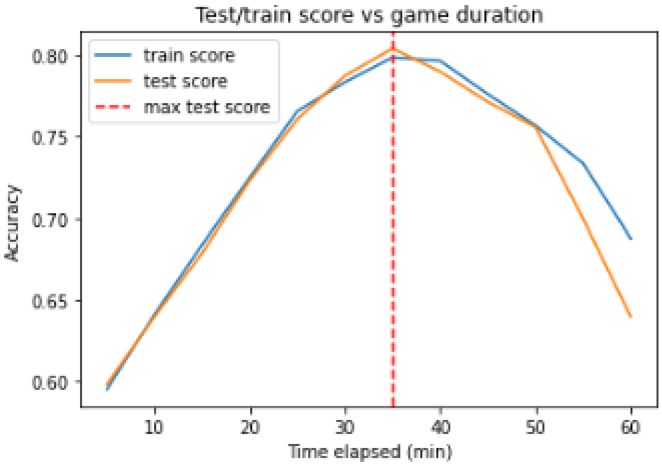


Figure 14: Variation of test/train accuracy as we vary the match duration

We observe that the team with the gold lead at any point during the match has a higher probability of winning the match. Of all the winning teams across the dataset, 98.4% of the teams won the match with more gold than the opponent. We trained a cross-validated logistic regression model using different durations of the match, to test prediction capability of the feature while the match is in progress.

- **Barracks kill**

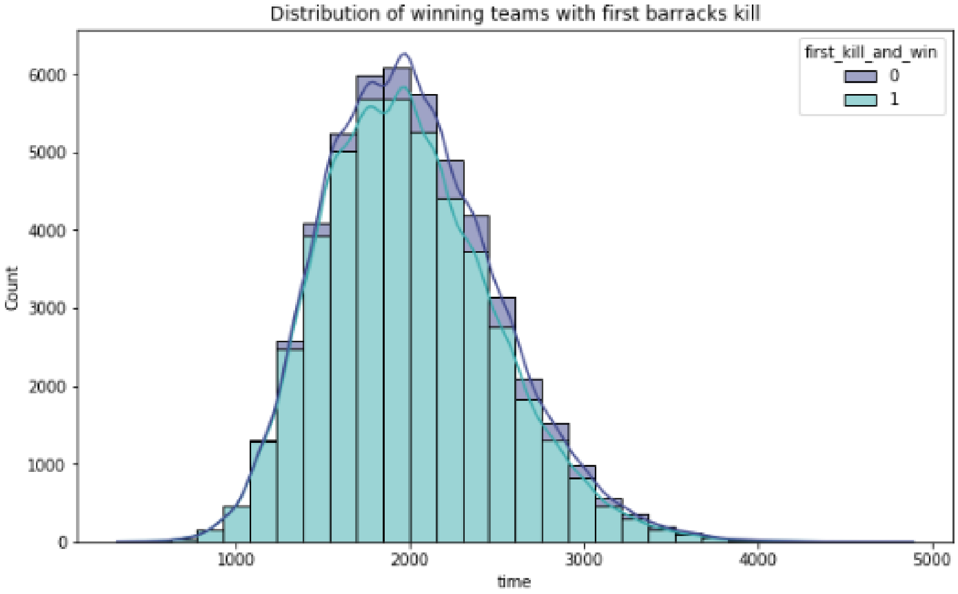


Figure 15: Winning and losing teams with the first barracks kill

By adding the name of the team with the first barracks kill at any point in the match to the dataset, we were able to observe that the values for this flag and the win flag were correlated. To identify the strength of this correlation, we plotted the distribution of the number of winning teams against the time when the first barracks were destroyed. We also stacked the number of times the team got the first barracks kill but still lost.

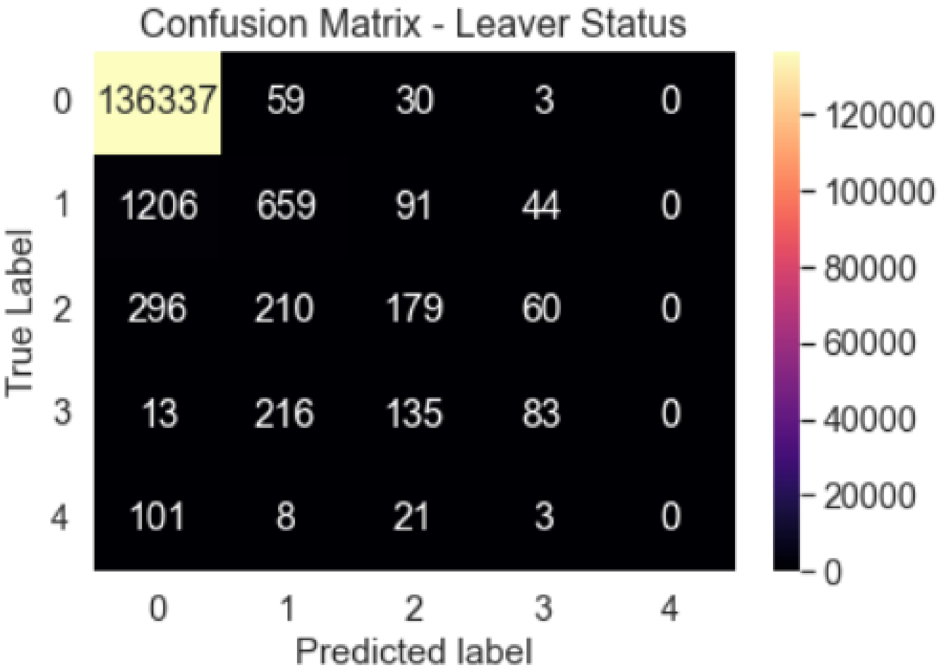
Leaving Players

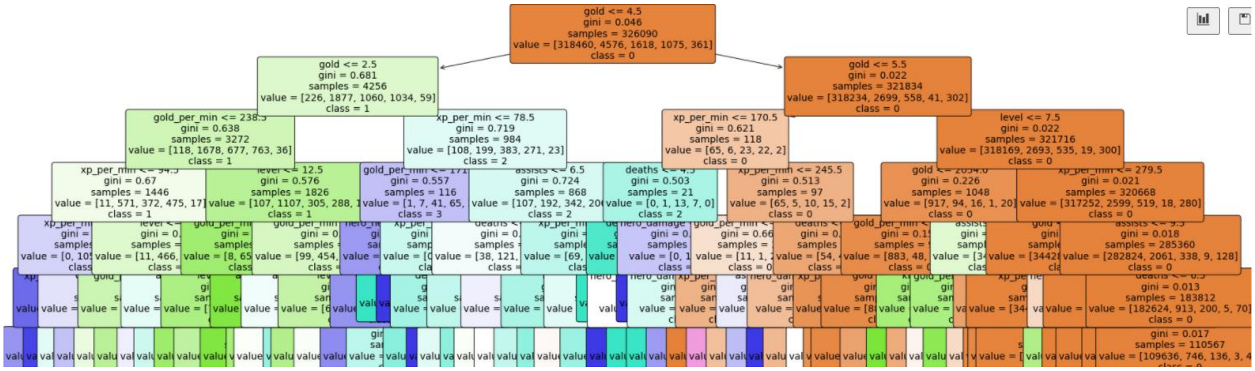
Using the matches dataset, we created 3 separate models to predict players leaving, a Decision Tree and Logistic Regression to predict if a player would abandon the game, and a Decision Tree to predict the leave status of the players.

Leaver Status

Decision Tree

Our full decision tree has a depth of 50 and 9000 leaves, after pruning we are able to reduce that down to a depth of 6 with 60 leaves. This final tree has a training score of 0.98320 and a test score of 0.98214. Below is the associated confusion matrix and decision tree.

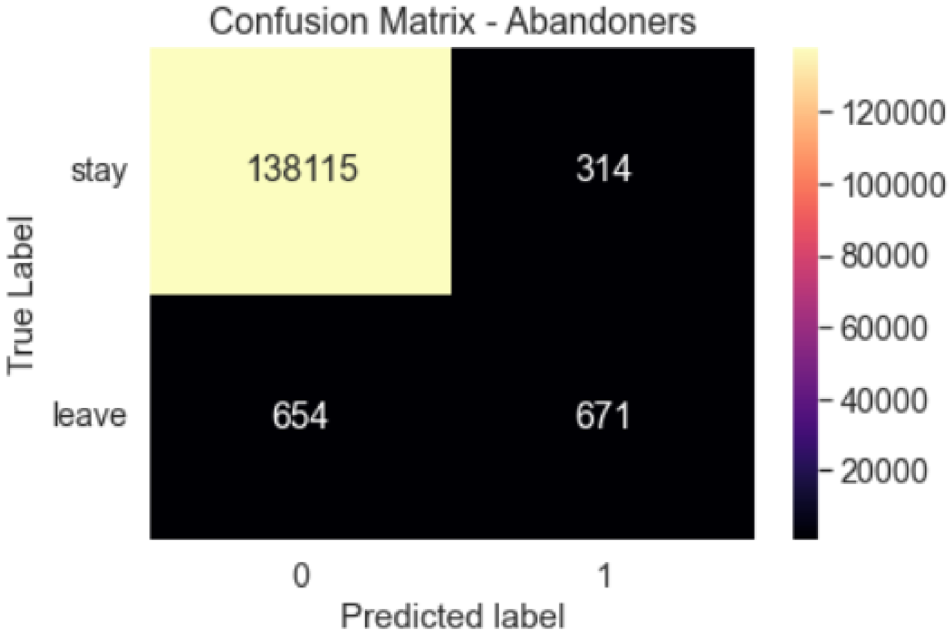


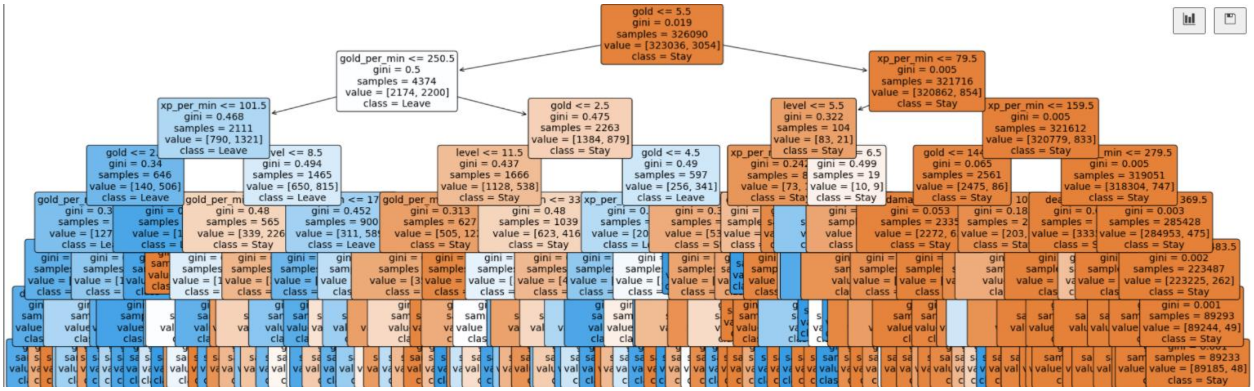


Abandonment

Decision Tree

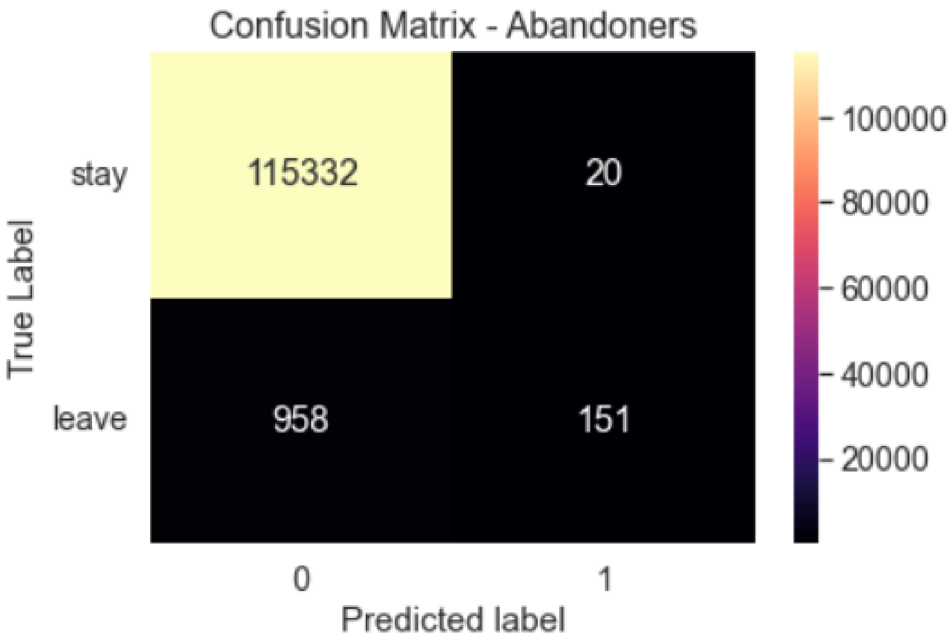
Our full decision tree has a depth of 34 and 2971 leaves, after pruning we were able to reduce that down to a depth of 7 with 105 leaves. This final tree has a training score of 0.99371 and a test score of 0.99307. Below is the associated confusion matrix and decision tree.



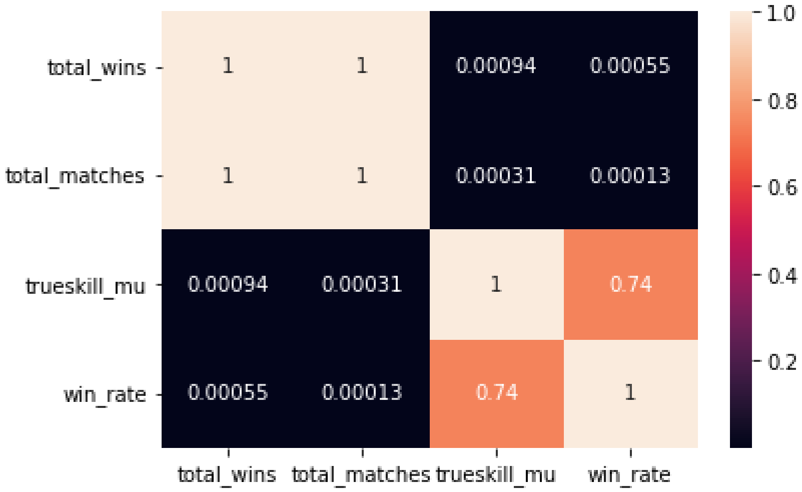


Logistic Regression

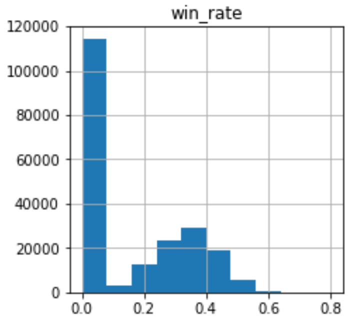
For our logistic regression we first used a standard scaler before training the model. Our training score is 0.99165 and our test score is 0.99160.

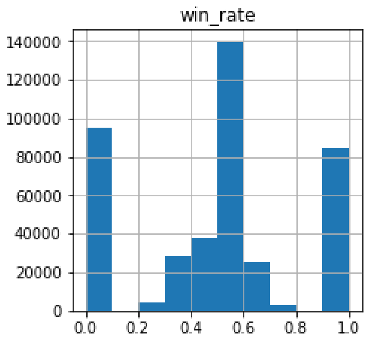


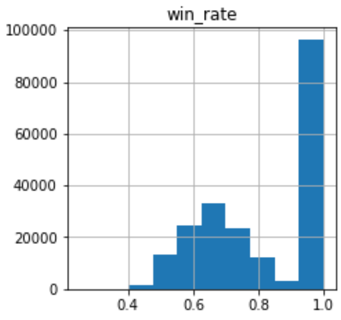
Player Skill Prediction



Correlation heatmap



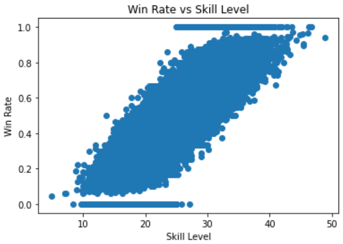


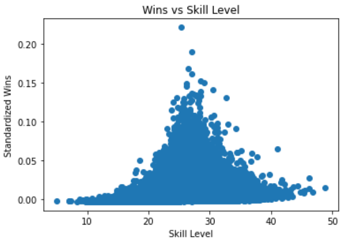


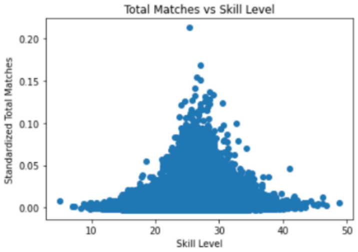
Beginner win rate histogram

Intermediate win rate histogram

Pro win rate histogram







Win Rate vs Skill Level plot Wins vs Skill Level Plot Total Matches vs Skill Level Plot

We started out by doing data exploration of the data set, creating plots and correlation heatmaps. It was found that the win rate has a strong correlation with the trueskill_mu. So we decided to look at the win rate for beginner, intermediate, and pro skill levels individually. As expected, beginners tend to have a lower win rate, intermediate players were more balanced in terms of win rate, and pros had higher win rates. There is a clear positive linear correlation based on the Win Rate vs Skill level plot as well.

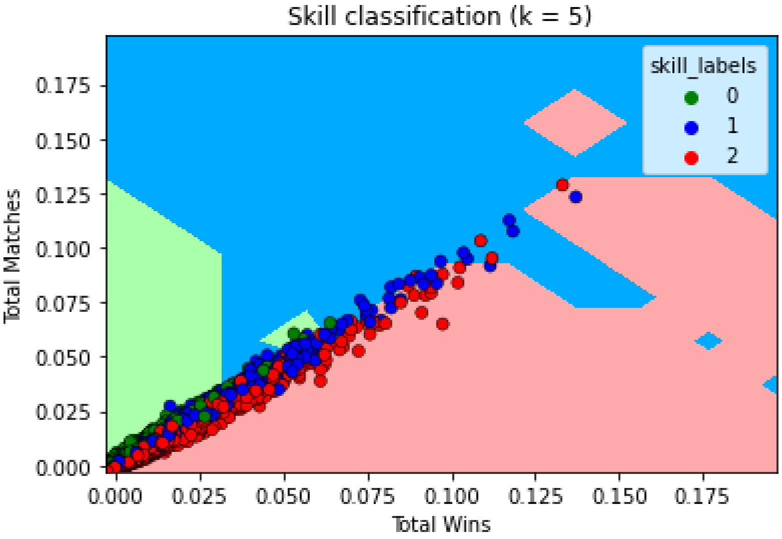


Figure 16: Graph of kNN model

Based on this initial data exploration, we decided to create a k Nearest Neighbors model using total wins and matches to predict the skill level, 0, 1, or 2. The model used a value of 5 for the number of neighbors. Due to the large size of the dataset, the model took a while to run, which did not allow for time to optimize the hyperparameter. We used an 80/20 split for training and testing data. The score on the test data was 0.698, which shows that the wins and matches features were decent predictors of skill level.

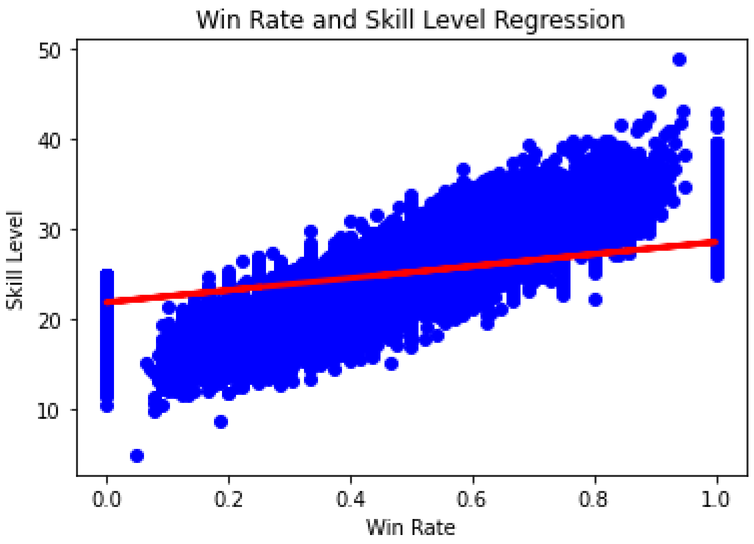


Figure 17: Graph showing linear regression of win rate and skill level

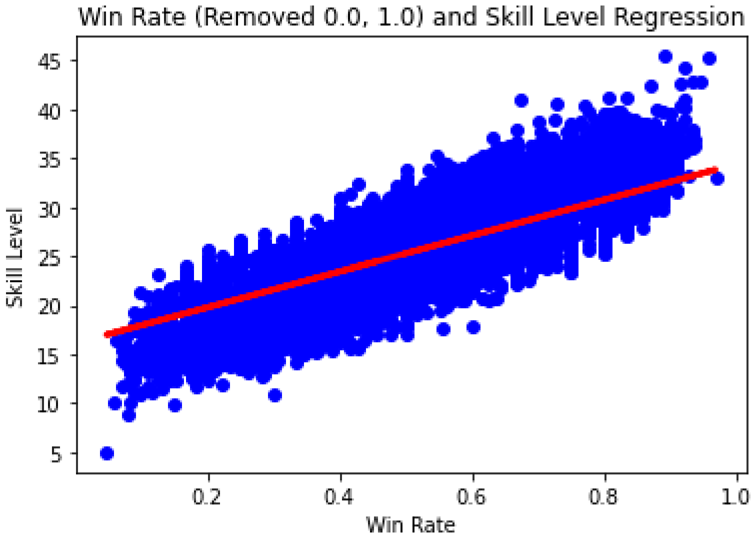


Figure 18:Graph showing linear regression of win rate without 0% and 100% instances and skill level

We wanted to explore other models to see if we could improve our accuracy of predicting skill. So, we looked into creating a linear regression model with the win rate feature. This time we used a 10-fold cross validation to train the model, with an 80/20 testing and training split. The model returned a mean score of 0.545, and when tested on the full test set, the MSE was 4.72 and an R-squared value of 0.547. This model turned out to be less accurate at predicting the skill of a player. However, we noticed that there are many players with 0%- and 100%-win rates, that have varying skill level. This is most likely because you can still gain skill level even while you lose games. So we decided to remove the instances that have 0% and 100% win rates and run another linear regression model to see if we can more accurately predict the win rate inside those win rates. As a result, the model performance increased, with a mean score of 0.678, MSE of 3.52 and R-squared score of 0.678. While it did outperform the other linear regression model, it performed about the same as the k Nearest Neighbors model.

Discussion

Professional hero picks

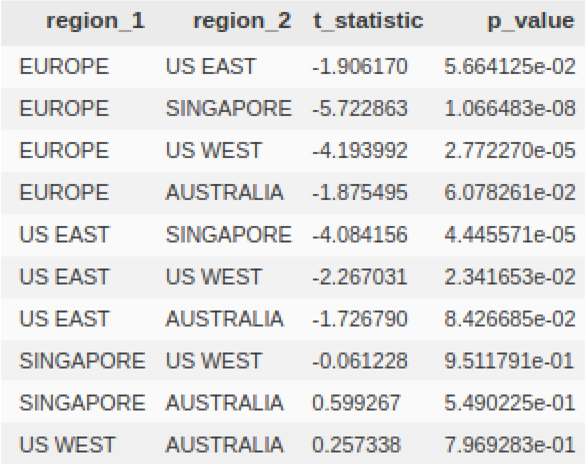
The top 20 plot for the most picked heroes gives us an insight into the perceived strength of the hero. When we compare this popularity rating of the hero to its win rate, i.e., the number of times a team won with that hero in its roster, we see that it negatively correlates to the win rate. Based on this, we can infer that the player perception does not necessarily translate to a higher winning chance for the team.

We also get the team compositions with the highest wins to get an optimal hero combination that maximizes the winning chances. We also trained a classifier on the team hero selection to predict the win. However, on further analysis we see that due to the wide selection of heroes, there are a vast number of unique combinations that teams can have. This results in a lot of teams having similar win rates. Subsequently, we get poor accuracy on the prediction model. Based on this, we can infer that the team composition is not a definitive metric that correlates to the win. We can also say that the heroes in Dota 2 are well balanced, where we can use multiple heroes to fill in the same role, and a combination of over-powered heroes is highly unlikely.

Association Analysis

Our support across the board for both matches and team selection is surprisingly low, and the confidence is also low though there is a clear difference with match rules having significantly higher scores here. With over 100 heroes in 5 different ways (10 including both teams), it can be difficult to create strong rules when the pairings can be sparse. With more data or clearer ways to bin the heroes we could dive deeper into the hero composition.

Game duration across region



region_1	region_2	t_statistic	p_value
EUROPE	US EAST	-1.906170	5.664125e-02
EUROPE	SINGAPORE	-5.722863	1.066483e-08
EUROPE	US WEST	-4.193992	2.772270e-05
EUROPE	AUSTRALIA	-1.875495	6.078261e-02
US EAST	SINGAPORE	-4.084156	4.445571e-05
US EAST	US WEST	-2.267031	2.341653e-02
US EAST	AUSTRALIA	-1.726790	8.426685e-02
SINGAPORE	US WEST	-0.061228	9.511791e-01
SINGAPORE	AUSTRALIA	0.599267	5.490225e-01
US WEST	AUSTRALIA	0.257338	7.969283e-01

The results of the pair-wise t-test are summarized in the table above. We see that the most statistically significant variance is for regions Singapore, US West and Australia. The game durations in these regions are longer compared to US East and Europe. Based on this observation, we can conjecture that the players in US East and Europe opt for a faster, more aggressive approach compared to the other 3 regions. This could also mean that the players in these regions are at a higher level, so the matches play out faster.

Predicting match winner

All features

Using most features to train a classifier, we observe that it is not a hard problem to model once the game ends. A lot of features are really good at predicting a win. The classifiers trained on this dataset get perfect train and test scores, which points to some features in the dataset having unit correlation with the win.

Economy

We observe that the gold level of a team, and the team in gold lead is a strong predictor. We got the highest training accuracy of 79.8% and test score of 80.4% after 35 minutes of gameplay. This is mostly because Dota 2 rewards players with gold every time they damage the enemy team's players or economy. An interesting observation is that even 20 minutes into a game, we can make a reasonably accurate prediction about the outcome by looking at the team in the gold lead.

Barracks kill

Based on the results and the distribution plot, we can infer that the team to secure the first barracks kill has a remarkably high chance of winning. This intuitively makes sense looking at the importance of barracks in a match. We also see that as time passes in the game, the correlation falls off to an extent. Any team that secures the first barracks kill in the first half of the game further maximizes its chances of winning.

Player Skill Prediction

Based on the models we created using the player ratings data set, it is safe to say that the win rate is a strong predictor of a player's skill level. It would have been interesting to see how the number of neighbors hyperparameter in the kNN model would have affected the accuracy. However, running a GridSearchCV with multiple values for number of neighbors would have taken multiple hours due to the size of the data set. Another interesting observation is that there is a wide range of player skill for win rates at 0% and 100%. This made it harder to predict the skill of players with those win rates. Inside those win rates, our linear regression model was much better at predicting skill.

Leaving Players

Our models show good results on predicting both abandoning players and leaver status. While these results are good, there are a lot of false positives and negatives especially since our datasets are so imbalanced resulting in most of the classes falling into the majority class of staying. We are also missing the 5 and 6 classes, NEVER_CONNECTED, which could be an issue with the data collection and would most likely help to reduce the false positive rates of leaving.

Future Work

Currently our model leverages the hero pick rate, gold, and barrack kills to predict the outcome of the match. On addition to that, we can add item purchase during the game and analyze its contributions towards the outcome of the match. Some of our hero analysis could be continued by taking their stats, roles, and lanes into account to make predictions on use and potential.

There are many directions that we can further this work, for both DOTA2 as well as extending into other MOBAs like League of Legends as well.

References

1. Real-time eSports Match Result Prediction by Yifan Yang Tian Qin Yu-Heng Lei from the Language Technologies Institute, Carnegie Mellon University
2. Machine learning models for Dota 2 outcome prediction by Kodirjon Akhmedov and Anh Huy Phan
3. Dota 2 prize pool information was collected from this website
4. The information about the game was collected from this website