CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Analysis of consumer behaviour among DotA 2 players and its effect on individual performance

Bachelor's thesis

Author: Vilém Krejcar
Study program: Economics and Finance
Supervisor: prof. PhDr. Ladislav Krištoufek, Ph.D.
Year of defense: 2021

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, July 26, 2021

Vilem Krejcar

Abstract

This thesis examine the possible effect of consumer behaviour, specifically gold sources distribution and effectiveness in spending, on rank between DotA 2 players. Relevant literature from the past years was summarised and dataset containing more than 91 thousand data points from OpenDota API used to determine the relationship was constructed. Ordered Logistic Regression and K-Means Clustering algorithms were applied on four datasets of 12 variables divided based on the player's position and the outcome of the game, concluding that only a fraction of the variation (McFadden's R^2 of 6.2% and Cox & Snell's R^2 11.7% on average) has been explained and cluster analysis resulted in a mean accuracy of 38.25%. Those results were supported by the correlation analysis indicating low correlation coefficient values of dependent variable *rank*. Despite such output, patterns were discovered across individual rank clusters from metrics acclaiming higher effectiveness of more experienced players regarding consumer behaviour with strong statistical significance. It has been concluded that the effect is present, yet its magnitude by itself is not sufficient to predict the regressand. The outcome implies possible presence of other variables that have an effect on the explained variable which are not included in the analysis.

JEL Classification	C55, C57, D90
Keywords	DotA 2, consumer behaviour, data analysis, Python, data mining, artificial intelligence, machine learning, esport
Title	Analysis of consumer behaviour among DotA 2 players and its effect on individual performance

Author's e-mail	18667904@fsv.cuni.cz
Supervisor's e-mail	ladislav.kristoufek@fsv.cuni.cz

Abstrakt

Tato bakalářská práce zkoumá možný vliv spotřebního chování, konkrétně distribuce zdrojů zlata a efektivity během nakupování, hráčů hry DotA 2 na jejich výkonnosti. Byla shrnuta relevantní literatura z posledních let a byl sestaven datový soubor obsahující více než 91 tisíc datových bodů z rozhraní OpenDota API, jež byl použit ke zkoumání tohoto vztahu. Na čtyři datové sady obsahující 12 proměnných rozdělených na základě pozice hráče a výsledku hry byly použity algoritmy uspořádané logistické regrese a K-Means algoritmus se závěrem, že byl vysvětlen pouze zlomek variability (průměrné hodnoty McFaddenova R^2 6,2% a Cox & Snellova R^2 11,7%) a shluková analýza s přesností 38,25%. Tyto výsledky byly podpořeny korelační analýzou, která ukázala nízké hodnoty korelačního koeficientu závislé proměnné *výkonnost*. Navzdory těmto výsledkům byly v jednotlivých výkonnostních skupinách objeveny zákonitosti potvrzující vyšší efektivitu zkušenějších hráčů v oblasti spotřebního chování se silnou statistickou významností. Závěrem byla přítomnost tohoto efektu, avšak jeho velikost sama o sobě není dostatečná k předpovědi závislé proměnné. Výsledek naznačuje možnou přítomnost dalších proměnných, které mají vliv na vysvětlovanou proměnnou, jež nejsou v analýze zahrnuty.

Klasifikace JEL	C55, C57, D90
Klíčová slova	DotA 2, spotřební chování, datová analýza, Python, dolování dat, umělá inteligence, strojové učení, esport
Název práce	Analýza spotřebního chování hráčů DotA 2 a jeho vliv na individuální výkon
E-mail autora	`18667904@fsv.cuni.cz`
E-mail vedoucího práce	`ladislav.kristoufek@fsv.cuni.cz`

Acknowledgments

I am inconceivably thankful to prof. PhDr. Ladislav Krištoufek, Ph.D., the thesis supervisor, who has provided me with valuable feedback throughout the creation of the thesis and has been directing my ideas to a more coherent structure. He has a tremendous overlap of knowledge in the field of econometrics, and his remarks have helped me to better understand numerous topics regarding the subject. It was my great pleasure to cooperate with him. Moreover, I wish to express my gratitude towards my dear colleagues and friends Bc. Petr Čala and Bc. Daniel Bartušek not only for their tips which stimulated ideas later implemented in the text, yet in academic life generally.

Typeset in FSV LaTeX template with many thanks to prof. Zuzana Havrankova and prof. Tomas Havranek of Institute of Economic Studies, Faculty of Social Sciences, Charles University.

Bibliographic Record

Krejcar, Vilem: *Analysis of consumer behaviour among DotA 2 players and its effect on individual performance.* Bachelor's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2021, pages 56. Advisor: prof. PhDr. Ladislav Krištoufek, Ph.D.

Contents

List of Tables

List of Figures

Acronyms

DotA Defense of the Ancients

DotA 2 Defense of the Ancients 2

MOBA Multiplayer Online Battle Arena

TI The International

LoL League of Legends

MMR Matchmaking rating

HP Health points

GPM Gold per minute

XPM Experience per minute

APM Actions per minute

AoE Area of Effect

ML Machine Learning

AI Artificial Intelligence

RL Reinforcement Learning

ANN Artificial Neural Networks

DT Decision Tree

API Application Programming Interface

JSON JavaScript Object Notation

CDF Cumulative Distribution Function

API Application Programming Interface

OLS Ordinary Least Squares

MLE Maximum likelikood estimation

Chapter 1

Introduction

Computer games have been on the popularity rise for several years, creating a solid and passionate fanbase. The popularity was so immense that conversations regarding moving e-sport (electronic sports) on the same level as regular sports, even inserting the most popular titles into the Olympic framework, have been rising. The first step has already begun, with a nudge caused by coronavirus - an official statement of the International Olympic Comittee[1] describing the creation of *Olympic Virtual Series*, an online alternative to the Olympic games consisting of 5 sports - rowing, sailing, cycling, baseball and motorsport - which will be supported by gaming equipment in a virtual environment.

From the very basic of things, everybody desires to improve their skillsets. Not only amateurs tend to look up to professional DotA 2 players who show off their strong mechanical skills, such as perfectly timed combinations of heroes' abilities, while the commentator is emotionally shouting how precise this action was. Consequently, there is a sizeable market where more experienced players coach the less skilled ones and provide tutoring in ability-based skills inside the game. However, above mentioned is only one side of the coin. We hypothesise that people may be putting potential significant metrics away, leading to other ways of improving. A significant volume of data is being collected every match depicting both teams' and individual players' behaviour. We would like to discover if players behave differently across rank clusters from the economist's perspective, such as the structure of their in-game gold sources and how they act as consumers. The conclusion from such analysis could lead to a list of possible predictors influencing rank with a respective volume, resulting in possible

[1]Source: `https://bit.ly/3eTMhrW`.

suggestions towards the community on improving one's playstyle not solely by casting heroes' abilities or positioning, yet by changing habits connected to income, spending and other related in-game actions.

The other parts of the thesis are structured as follows: Chapter 2 should build up a fundamental overview as a necessity for the topic understanding consisting of both DotA 2 in broader e-sport context with market competition comparison plus its history, with relevant in-game design and DotA 2 mechanics. Chapter 3 explains how the dataset was gathered collectively with rationale behind and methodologies implemented, followed by closer overview of the individual variables. Chapter 4 will further depict mined data from the look of an econometrician and will build up theory as a prerequisite for algorithms used and estimation, following in Chapter 5. Finally, Chapter 6 describes the summary of the analysis output and Chapter 7 concludes the research.

Chapter 2

The effect of consumer behaviour on the rank of DotA 2 players

2.1 Theoretical overview

2.1.1 Brief history of DotA 2

Defence of the Ancients 2, or shortly DotA 2, is a free online multiplayer game released by Valve on July 9^{th} 2013[1], which emerged in one of the most flourishing Multiplayer Online Battle Arena (MOBA) games on the planet. Its ancestor, Defense of the Ancients (DotA), was integrated into a game developed by Blizzard Entertainment, *Warcraft 3: Reign of Chaos*, in 2002 as a custom map expanding the players' possibilities on top of the regular game experience. It has quickly gained popularity across the Warcraft 3 community, and its player base has advanced significantly. As a result of the long-term hard work, Valve, a major company in the gaming industry, has invested in this community project, and a version of today's tremendously popular game has been born.

Ever since its releasement, both DotA and later DotA 2 had a massive upward swing in active player volume and widespread recognition. This factor got reflected in the competitive scene - tournaments and competitions started relatively modestly with symbolic prices such as refreshments or items of minuscule value. As the game's popularity exploded, so did the tournaments. Minor prizes have been growing and soon enough, specifically in 2011, the first enormous DotA 2 tournament has taken place.[2] With historically the first

[1]Source:`http://www.valvesoftware.com/`.

[2]Source: `https://liquipedia.net/dota2/The_International`.

e-sport tournament featuring a prize pool exceeding one million dollars, the tradition of The International (TI) tournaments began. The first TI took place in Cologne, Germany and distributed prices worth approximately 1.6 million Euro, way above any other e-sport tournament at the time. It still remains among the fifty most generous tournaments in all of the e-sports. After ten years of continuous progress and venues distributed across three continents, TI in 2021 will be held in Europe for the second time since 2011 in Bucharest, Romania with an estimated prize pool of more than 40 million dollars. This is a 6 million dollars increase from the last TI, making a prize pool gap between DotA 2 and other games more than 25 million dollars.[3]

2.1.2 Elementary game design

As mentioned above, DotA 2 is a multiplayer game. Two teams consisting each of five players are competing at a time, plus an option to have a spectating coach is present. Both teams are clearly identified and have their specific names, Radiant and Dire.

A DotA 2 match has a unique map with Radiant and Dire bases located in the bottom left corner and upper right corner, respectively. The base consists of a place in which the players are beginning the game and where they appear after their death, so-called *spawn points* and further infrastructure - multiple towers, three sets of barracks, a shop and the most important building in the game, an *Ancient*, also a reference to the game's title. Also, multiple smaller buildings which have close to no impact on the game are located there. Outside of the base, there are two jungle areas separated by a river in the middle, a Roshan's pit (more detailed information will be presented in Section 2.1.5), two secret shops and outer towers. All of above described can be seen in Figure 2.1.

The goal is to destroy the opponent's Ancient. There is no other way of ending the game apart from all team players leaving the game.

2.1.3 Pre-game & hero theory

Every game starts with matchmaking. During matchmaking, players are randomly assigned to teams based on region, Matchmaking rating (MMR) and

[3]Source: `https://www.esportsearnings.com/tournaments`.

Figure 2.1: DotA 2 map as of patch 7.29d

game mode. Regions could be selected manually by an individual, while MMR being precise and reflects on the player's rank obtained from preceding performance. For our particular study, we will restrict the set of games modes to two, *Ranked All Pick* for public matches and *Captains Mode* used in competitive gaming.

After players are allotted into the game, the draft phase is initiated. Drafting does differ across the two game modes, both in the picking order and the way hero pool is restricted, nicknamed *banning*.

Ranked All Pick allows all ten players to vote heroes for bans individually during the first fifteen seconds of a draft and after the countdown, half the heroes nominated for bans will be selected at random and banned. From the remaining hero pool, heroes are selected according to Table 2.1.

Table 2.1: Drafting scheme for Radiant All Pick as of patch 7.29d

Radiant	**Dire**	**Dire**	**Radiant**	**Radiant**	**Dire**	**Dire**	**Radiant**	**Radiant**	**Dire**

Captains Mode allows each team to precisely restrict a pool of 7 heroes and bans are distributed as follows: two initial bans, two picks, three bans, two picks, two bans and a final pick, supposing that each banning or picking is done twice by teams taking respective turns.

From the remaining pool, each team is obliged to select five heroes to play with. The hero pool of DotA 2 is 121 heroes as of patch 7.29d, and heroes are divided into numerous categories. To begin, every hero character has Health points (HP) and mana. HP bar represents a life of a hero. In the event of HP

being equal to zero, the hero dies. Mana is used as a mean of casting spells. Both HP and mana are being regenerated every second, increasingly with levels.

It is crucial to group heroes by their role during the match. Based on this, we divide heroes into two groups - carry and support. Carry type is vulnerable in the early game and is supposed to scale with items purchased and levels gained, potentially game-winning given the right circumstances. On the contrary, the goal of support players is, from their very name, to aid carry players and ensure they are getting space necessitated.

2.1.4 Match description

Following the drafting phase, the game itself begins. Players are allocated to three lanes based on their classification from the previous paragraph and team consensus. Usually, the distribution is following: there is a one-on-one carry matchup on the middle lane, abbreviated to *mid* in DotA 2 jargon. Top and bottom lanes are both occupied by four contestants, one carry and support from every team. It can happen in several strategies that one hero is left utterly alone versus two or even three opponents, and extra attention is furnished to the other carry to ensure easier start. Such lane with two supports and a carry is termed *trilane*.

Apart from this, heroes are assigned specific positions depending on their scalability and need to obtain gold, from position one being a carry with the highest priority; to position five being one of the two supports. Additionally, we can define hard and easy lanes. For Radiant an easy lane is bottom and a hard lane is top, vice versa for Dire. Thus, there is a conflict between Radiant easy lane and Dire hard lane on the bottom lane. Bottom lane typically consists of Radiant positions one and five versus Dire positions three and four. Carry player being on the hard lane, position three, is called *an offlaner*. Both teams' position two players commence the game on mid lane, seizing the place to showcase their skills in a duel.

After several minutes after the initial horn, recently defined theoretical boundaries are blurring and game variability increases. Either supports are making regroupings to mid lane in order to create pressure on the opponent mid player, or mid player comes to one of the side lanes searching for kills. The middle

stage of the game usually contains numerous team fights and rotations, while position one players are gathering more items to fulfill their potential, rising on power. Teams are trying to create pressure on and possibly destroy towers, resulting in the accumulation of map presence and pressure on the opponent. With more tower destruction, a team can reach opponents' base, being the main target in order to win. After three outer towers, barracks have to be besieged. Succeeding demolition of barracks, more powerful units are spawning, nicknamed *super creeps*, and after clearance of all three sets of barracks, very powerful *mega creeps* are spawned. While getting to the Ancients, only one set of barracks and the last two towers in front of Ancient ought to be destroyed. With falling Ancient, the game ends.

2.1.5 Gold and experience

Concerning the matter of the analysis, it is essential to define what monetary decisions are players allowed to make by the game environment and what are different sources of gold and experience.

As the match starts, every player is endowed with six hundred gold. The initial purchase is crucial for the first several minutes of the game, wrong decisions can lead to a deterioration of lane equilibrium during the early game, decreasing the team's chance to win. There is also a passive gold increment for every player, increasing from 100 to 128 GPM with the time trend.

Apart from passive gold, players make gold in numerous active fashions. The most usual way is nicknamed *farming* or *last hitting*. On a minute basis, every team's lane is accompanied by NPCs (non-player characters) termed *creeps*. A tiny volume of gold and experience is credited to a player if he can perform a last hit and kill the creep with his attack, creeps being the main source of mentioned two commodities. From Subsection 2.1.4 and respective positions, heroes are prioritised in last hitting. The role of the supports can be further extended as a responsibility to ensure safe farm for the carry.

Moreover, a player can hit his own creeps if the creep's HP drops below half. If he manages to do so, it is called *a deny*. Denied creep gives the opponent no gold and only a half of experience points, the other half allocated to the denying player, marking denying as a powerful tool. Denying can also be used

on towers under ten per cent of HP, giving the player portion of the gold reward, or on players with HP dropping below a quarter, depriving opponents of all the gold and experience.

It is not only lane creeps that are to be farmed. Both jungles are accompanied by neutral creeps, which any player can farm. Neutral units provide gold and experience, yet neutral items can also be dropped on their death after several time thresholds. Items generally will be described in following Subsection 2.1.6. Further, Roshan has his pit in the river between mid and top lanes, as shown in Figure 2.2, being the most potent neutral unit on the map. It is laborious to kill Roshan, and usually more than two heroes during mid-game ought to do it. Nevertheless, the rewards are significant - on his death, Roshan drops *Aegis of the Immortal*, granting a second life to its owner, disappearing after the resurrection. On second and further Roshan deaths, the kills rewards are increasing.

There are plenty of other ways to seize gold, but we will not disclose them further due to their marginal effect and reader's comfort. However, the concept of runes is worth mentioning from numerous perspectives. Runes are objects that are thought to spawn in specific intervals at respective places. There are two main categories of runes, *bounty* and *power*. With a start of the game, bounty runes spawn in both teams' jungles. Since then, they spawn every two minutes, starting with the third minute. Bounty runes supply a lump sum reward for the team that picks it up, increasing with every new spawn. Power runes, spawning in the river, provide their picker with extraordinary power such as invisibility, fast movement speed, double damage, regeneration, or creation of two illusions. Power runes appear first at six minutes and appear every two minutes, randomly at one of two possible spots with effect chosen randomly from mentioned above. Graphical illustration of rune spawn points is shown in Figure 2.2.

2.1.6 Items in DotA 2

Another critical feature is itemisation. By buying items, heroes obtain substantially more attributes than solely levelling up and many beneficial effects are conferred to a hero. In general, there are two types of purchasable items. One of which are consumables or items meant to be used once, disappearing

Figure 2.2: DotA 2 minimap with bounty rune places marked yellow, power runes marked purple and Roshan pit in a red circle.

afterwards. They generally do not have many passive impacts, but the active effects are strongly beneficial, especially in the early game. For example, HP or mana regenerating items such as *Healing salve, Clarity* or *Bottle* and items that are intended for transportation, namely *Town Portal Scroll*. Aside from consumables forming a miniature ratio, there is a wide spectrum of items meant to be combined and upgraded into more powerful pieces. Those items could include either active or passive effect, yet an effect that alters the character's surroundings described as Area of Effect (AoE).

2.2 Previous findings

While analysing the historical volume of DotA 2 active players from *Steam Charts*[4], it is possible to spot a possible correlation of its peak during 2016 with the volume of academic research regarding DotA 2. Mora-Cantallops & Ángel Sicilia (2018) have provided an exhaustive literature review in the field of MOBA games until the year 2018, concluding that researches focus mainly on League of Legends (LoL) and DotA 2 - together with DotA forming more than 95% of listed scientific MOBA literature.

[4]Source: `https://steamcharts.com/app/570`.

A wide range of topics related to DotA 2 has been explored. Leaving aside direct effects on team success and performance, we can find articles explaining social aspects of MOBA games, such as communication via chat or voice chat and direct connection to toxicity[5] of players from Kwak et al. (2015) where the reporting system in LoL was analysed and in-game conflicts were depicted. Moreover, Tseng (2011) has used one's perception and motivation for playing the game, such as the pursuit of winning or tendency to relax, to cluster online game players and their respective consumer habits. The study revealed that players whose main aim was to defeat the game, so-called *aggresive players*, are more likely to perform a greater volume of game-related money transactions. Further, a theory of the match as a whole, specifically minor unbalances and uncertainties occurring throughout an individual match, is narrated by Simon (2014), giving a detailed analytical report of consequences caused by the game's progress together with its velocity. Finally, Winn (2015) investigates design and mechanics of DotA 2 and LoL and the extent to which both influence dramatics for the spectators. The author uses terms *uncertainity* and *inevitability* as a necessary condition for a drama to exist and tries to explain their variation caused by particular events throughout the game.

There is also a considerable amount of research concentrating on the investigation of success factors in MOBA titles. Rümmele et al. (2013) writes about the consequences of diverse team compositions and how they contribute to a team's winning odds. The study concludes that having two or more friends on the same team increments to the probability of a possible victory, similarly with increasing the team's expertise, clearly defining individual roles within the team, selecting proper leaders and picking heroes well-known by the team members. The importance of mechanical skills and metagaming is broken down by Donaldson (2017). From the study's findings, we can infer that for the most significant impact, it is not sufficient to only master one or the other; both knowledge of a current patch and mechanics such as timings, positioning, and ability usage have to be obtained. Yang et al. (2014) conducted an interesting study of DotA 2 fight patterns via graph metrics and decision trees to extract rules leading to optimal fight tactics, conditioning to win. For the early stages of the game, models produced a set of commands that should result in a higher probability of winning. On the other hand, for a late-game stage, no patterns leading to an advantage were found. The aforementioned is rational because as

[5]Inappropriate behaviour towards other players.

the game progresses, it is probable that the gap between the teams' likelihoods to win is sizeable, and thus a single fight has only a slight effect on the game's outcome. Effect of rank on spatio-temporal[6] changes in DotA 2 have been studied in Drachen et al. (2014) with statistically significant results. Authors have found out that with increasing ranks, the inter-zone movement has increased, and the average team distance, defined as an average Euclidean distance between individual players during a single match, has been lower for rank-wise better players. Findings may be interpreted as follows: more skilled teams tend to stick closer together and overall have a higher position variation, meaning they use more space on the map while sticking closer together. This results in higher pressure on the opponent and a higher backup chance if an enemy team decide to outnumber them. Behaviour differences are also examined by Castaneda et al. (2016). The study differs by using a qualitative questionnaire for tested players to support analytical research with an eye-tracker collected data. We can observe that with increasing rank, players tend to spend less time looking at the shop button on-screen and are more aware of health points and mana bar. Following research of Xia et al. (2019) focusing on predictors of success between professional Chinese and American teams during TI 4, we can see multiple mentionings of Drachen et al. (2014) introduced above. The study has concluded that the most effective predictor of success on the professional scene is *kills by multiple players*, a sum of kills with more than one player as a damage source in the death log. This outcome also supports the arguments of Drachen et al. (2014) in the sense of team play importance.

Considering research formulated after previously mentioned Mora-Cantallops & Ángel Sicilia (2018), methodology and algorithms applied have altered with the growing popularity trend of Artificial Intelligence (AI) and Machine Learning (ML) implementations in academia generally. With increasing computing power, more advanced and capable algorithms such as Artificial Neural Networks (ANN) commenced to be used by researchers to discover patterns that standard statistical methods such as linear or probabilistic models are not capable of revealing. An example of such predictive modelling is Beskyd (2018) using ANN and Decision Tree (DT) models with results of an accuracy slightly higher than fifty per cent on test data for both algorithms. However, we can assume that author has been overfitting his ANN model since the lowest train sample accuracy was 85.3%. A comparison between the linear model and AI-

[6]Refers to space and time, respectively.

based model is made by Akhmedov & Phan (2021), resulting in an accuracy of 82% using Linear Regression compared to 88% of ANN and 93% of Long Short-Term Memory models[7]. Initiative to perform micro-predictions using Deep Neural Networks has been shown by Katona et al. (2019). Researchers used Deep Neural Networks to create a model prognosticating the probability of any player dying in the following five seconds, achieving a 37.7% precision with 72.5% recall. The authors were content with such result, considering death is a rare event occurring within 1% of the author's data. Also, a question of artificial players creation has been raised, Kočur (2018) programming a bot player with limited portfolio actions. This bot was afterwards compared with integrated DotA 2 artificial players, achieving higher efficiency in multiple factors and defeating them.

However, plausibly the most influential and practically implemented scientific paper from all of the above is OpenAI et al. (2019). Using Artificial Intelligence for a solution to the game via Reinforcement Learning (RL), AI has been trained by playing with itself for more than ten months. With numerous hyperparameter calibrations and model improvements, during the special event match at TI in 2019, the set of agents has successfully defeated the best team of that time, *OG*, twice in a row. To conclude, it has shown that even such complex mechanics as those implemented in DotA 2 can be dominated by AI using intelligent agents' analyses and predictions.

2.3 Our contribution

Following literature research, let us change the scope of writing and focus mainly on insights this study can produce. There is certainly untrivial number of research concerning DotA 2 and MOBA genre, as shown in Section 2.2. However, we can recognise that none of the above concentrated chiefly on consumer behaviour inside the game, considering it only as a side variable. We thus find space for a question that has not yet been answered, *is there enough evidence to prove that rank clusters are determined by consumer behaviour.*

By taking a closer look at the factors affecting in-game currency, gold, and its manipulation, if enough evidence is presented, we could derive different sources of improvement for the DotA 2 users. It would support motivation for players

[7]LSTM models are special examples of ANN.

not to consider earning and spending gold wholly as a byproduct of mechanicals rather than a skill that could be mastered and used as a future benefit. We suppose high rank could be driven up by the effectiveness of gold usage and the rationality of individual monetary decisions. As players obtain greater skill, we expect the microeconomic variables depicting the distribution of gold sources to focus less on fights and heroes, and pointing towards greater share of gold from creep last hits. Moreover, we expect higher ranked players to die less and be much more effective in spending.

Lastly, we may observe that preceding studies analyse games at a particular time in past. Because DotA 2 is developing with every patch, we expect that tendencies from several years may not be persisting until this point in time. Conducting new research, current trends could be discovered, and the academic community and player base are presented with up-to-date knowledge.

Chapter 3

Data mining and collection

3.1 Data extraction and dataset formation

From the statistics depicting the number of current active players mentioned in Section 2.2, we can deduce that for the past twelve months, as of June 2021, the average DotA 2 active player base was around 417 thousand players.[1] This number describes the average number of players with the DotA 2 game open at any time point during the last year. From this estimate's observation, we are convinced that there is a massive volume of games played every day, which could be used to answer the thesis question from Section 2.3.

To begin with, Python was chosen as a programming language to obtain the before-mentioned data in Jupyter notebooks for the sake of package availability and past programming experiences. Favourably, there are numerous ways to reach the desired data: an official Application Programming Interface (API) created by Valve is present together with an open-source[2] project called OpenDota with its API and many other platforms gathering information from DotA 2 matches, namely DotaBuff, GosuAI and STRATZ. In our case, OpenDota and its API was used for the data gathering. Following paragraphs will describe the process of data mining.

Firstly, an exploration of possible API requests provided us with a clear view of how to request and download the data required. One of the requests would list brief information about the last one hundred public games played. Next, we applied a filter that removed redundant noise connected to the whole

[1] Source: `https://steamcharts.com/app/570`.
[2] Created, modified and developed by the community.

spectrum of game modes and various game lengths provided from these one hundred games to find comparable games. We restricted the mentioned pool to games with a duration between circa sixteen and seventy minutes, game modes mentioned in Subsection 2.1.3 together with lobby types *ranked* and *tournament*. Filtering has left us with approximately 70 per cent of the 100 matches, and this particular task has been repeated multiple times with the aim of notable sample acquisition.

The filtering output comprising the specific information was delivered in JavaScript Object Notation (JSON) format, which was additionally manipulated and from which the dataset was derived. From every match, we extracted the *match_id* and sent another request to the API with a response containing detailed information about an individual match. Afterwards, from the match, we acquired data points from which we constructed our dataset. The structure of a single data point is divided into two categories - the first one includes information specific to the match, and the second one holds details of the player-specific actions throughout the game. The context and documentation of used variables will be given in the following Section 3.2. Finally, a function iterating through the dataset by data point additionally assemble the data into a pandas DataFrame, resulting in the final dataset.

We have accomplished the data mining part by doing the steps mentioned above, which played a significant role and was fairly time-consuming. We may now lead to the definitions of individual variables used, start exploring the properties of our dataset, and finally explain variation in the dataset by modelling it.

3.2 Variables description and the rationale behind

As mentioned in the last section, our data point is made up of match- and player-specific information. Let us start by naming variables that are used exclusively to filter and identify individual matches and players, *match_id*, *player_id*, *game_mode*, *lobby_type*, variable *patch* to have a sanity check regarding the uniformity of mechanics since the game changes across different patches, and finally *duration* denoting the length of the game in seconds.

To start with variables included in models, it is essential to define the independent variable, in our case *rank*. We designed this variable as a categorical variable, and our goal was to keep the number of levels as low as possible with possible future modifications. Each class has its respective sub-ranks: 1 being the lowest among the class and 5 the closest to the next rank. Only the highest one, *Immortal*, has no sub-ranks and is counted as a whole.

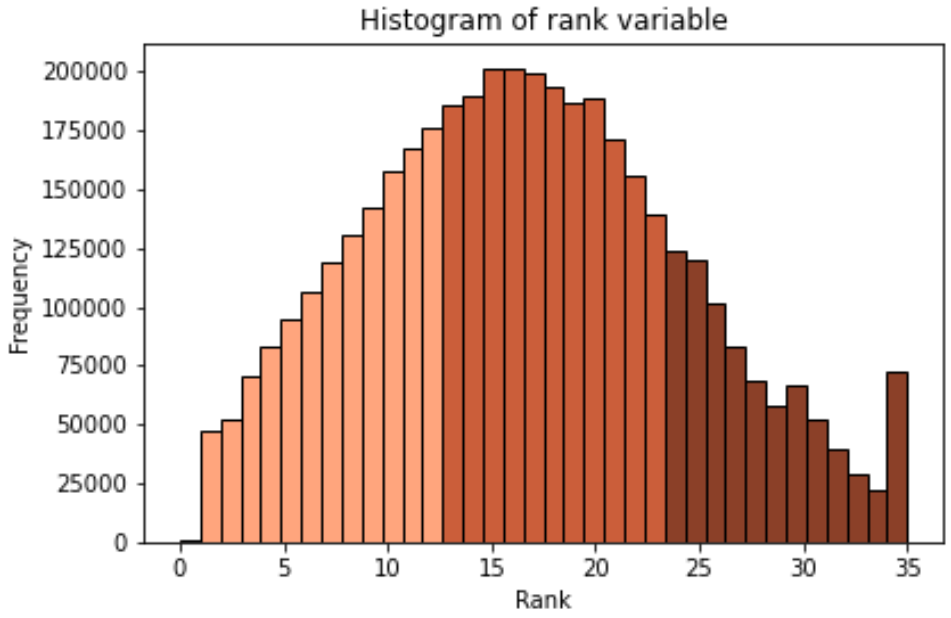

Figure 3.1: Distribution of DotA 2 rank tier as of June, divided into *Low*, *Mid* and *High* rank clusters from left to right respectively

For the time being, we decided to make the following clusters: *Low*, *Mid* and *High*. We defined following intervals for the respective levels, as can be seen in Figure 3.1:

$$\text{Low} = [\text{Herald 1}, \text{Crusader 3}),$$

$$\text{Mid} = [\text{Crusader 3}; \text{Legend 5}),$$

$$\text{High} = [\text{Legend 5}; \text{Immortal}].$$

Thanks to our belief that consumer behaviour can variate depending on the game result, we also added a boolean[3] variable *won*, which will be further used to divide data into won and lost games and used to evaluate the hypothesis.

[3]Having values of either zero or one, also termed a dummy variable.

Another dummy variable *is_radiant* does not differ from the previous one and is used as a data divider based on the player's fraction.

We also included several metrics that are used to compare players' individual performance such as *GPM*, defined as

$$GPM = \frac{\text{total gold earned}}{\text{duration}/60}$$

or simillar one, Experience per minute (XPM), defined as

$$XPM = \frac{\text{total experience gained}}{\text{duration}/60}$$

and overall measure of player's interactiveness and presence within the game, APM with following definition:

$$APM = \frac{\sum \text{actions}}{\text{duration}/60},$$

where by action we mean any mouse click on the game screen or keyboard shortcut press. Further, we defined a variable called *spent_gold* which follows formula

$$spent\_gold = \frac{\text{gold spent}}{\text{total gold earned}}$$

to measure the effectiveness of gold usage among players. We have two contradictory hypotheses connected to this variable, first stating that more proficient players would try to effectively spend every single gold and converge to one, the other one saying that players with a higher rank would save an amount of money for *buyback*, or the possibility to play money in exchange for an immediate resurrection. This mechanic is very crucial especially in the late game, and the cost of buyback is given by following formula:

$$buyback\ cost = \left\lfloor 200 + \frac{\text{gold earned}}{13} \right\rfloor,$$

where $\lfloor\ \rfloor$ symbolizes floor of a given number. Saving gold for buyback should, in our beliefs, shift the ideal *spent_gold* slightly above 0.9.

Following variable is connected to buying items, specifically consumables. Their specifics were explained before in Subsection 2.1.6. This variable is named *consumables_purchased* and is introduced as a total sum of consumable items

bought during the game. We believe that more mature players will tend to react more vitally to the current situation on the map and will thus have a higher volume of bought consumables, seeking the advantage from microdoses of HP and mana, allowing them to stay longer on the lane without any backing into the base. The hypothesis also supports that less skilled players tend to see only high-value items and upgrades, commonly forgetting smaller items and their non-marginal effects.

To reflect on the possible hero roles in the game (carry and support), we created a variable *is_carry*. The differences between those roles were mentioned in Subsection 2.1.3. It is commonly known that support will have lower GPM and XPM because they help carry players to become more powerful and leave all the last hits and farm to them. At first, the idea was to set the variable's values solely based on the picked hero. However, it is not uncommon that heroes meant as supports are also picked at carry positions, and thus we changed the classification. We took every team present in the data and divided the players based on GPM, resulting in two support and three carry players. We believe that division by GPM will result in more accurate classification since it is less probable that support heroes will gain more gold and experience than carry heroes.

Finally, there will be a bigger group of variables depicting the origins of gold and experience for players. A volume of individual sources will be defined for each possible source indicated as significant - in fact, there were other gold and experience sources, but we evaluated their effect as marginal and thus irrelevant. Those sources can be both positive or negative. For example, we have defined variables describing gold obtained or lost from death, spending on buyback, buildings, heroes, creeps, neutrals, Roshan and runes; and experience sources coming from heroes, creeps, Roshan and other factors combined, called others. With enough evidence provided, an ideal distribution will be discovered, leading to possible conclusions.

Chapter 4

Dataset properties and description

4.1 Basic properties of the dataset

After the menial process of data mining described in the previous chapter, final dataset has been gathered. We have mined more than 136 thousands matches, meaning more than a million and quarter individual data points have been obtained. Nevertheless, this number has to be lowered by a significant amount due to insufficient data quality. The cause of such a problem may be that OpenDota discards parts of the data after several days and keeps quite general information. After the data filtration, 91 860 data points have remained appropriate for the analysis.

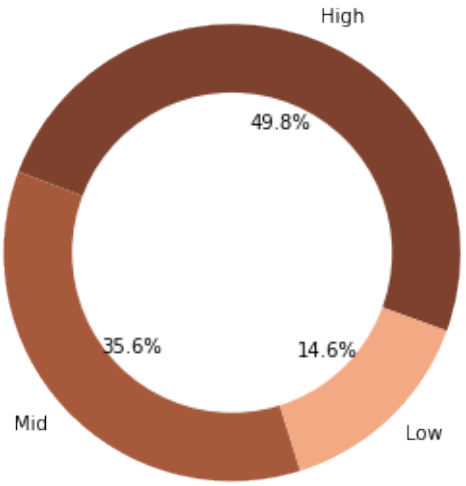

Figure 4.1: Rank distribution across the dataset

Observing the *rank* composition of the data, it can be seen that majority of the players is assigned to *High* cluster with nearly half of the data points

(49.8%), followed by *Mid* (35.6%) and closing the statistic with *Low* representing 14.6% of the observations, graphically in Figure 4.1. The *Low* rank gap is likely caused by the fact that players need to explicitly consent with full data extraction at the third party website in order to have complete data available online. We believe that some less-skilled players do not even know about this feature and thus do not grant access to their data. Further, this consent is a possible cause of lower data quality, as mentioned at the beginning of this subsection.

In order to give the reader a better insight regarding the distributions of predictors, we have decided to create a sequence of tables which plot the individual variables and portray their occurrences:

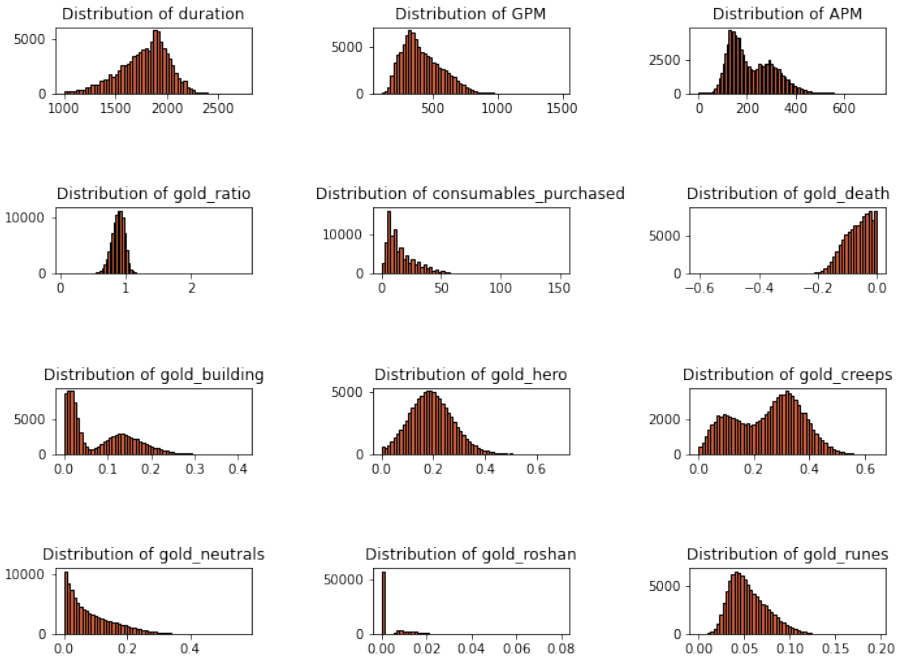

Figure 4.2: Histograms depicting distributions of independent variables included in the final model

It can be inferred that most of the regressors have quite well-behaved distributions, especially *gold_hero* seems to be reasonably similar to the bell shape curve, and thus is almost perfectly normally distributed. Other variables resemble a normal distribution, with different skewness and kurtosis coefficients, or having observations highly concentrated in tails. Some discrepancies arise from the definitions of the variables themselves - defined only positive, such as *gold_neutrals*, or strictly negative like *gold_death*. This is especially true for variable *gold_roshan*, which mode is equal to zero, having more than fifty thousand players not defeating Roshan during their game. Figure 4.3 is only an extension of Figure 4.2, and it depicts variable *gold_roshan* if Roshan has been defeated because the graph has been skewed by large volume of games with no Roshan kills.

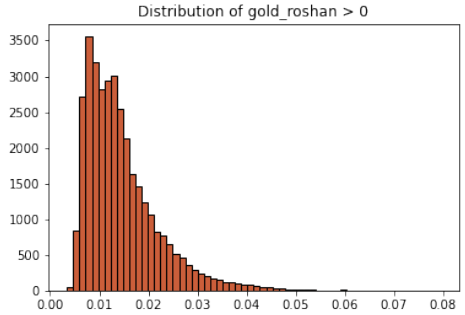

Figure 4.3: Spread of *gold_roshan* variable strictly greater than zero

Such findings will help the models to estimate the possible relationships more accurately as the regressors have almost no extreme values, and thus we expect the output to be more stable with lower standard errors.

4.2 Multicollinearity

A common problem across datasets is multicollinearity. From Wooldridge (2009), *multicollinearity* is a phenomenon describing if a specific value of one of the independent variables drives other variables' values in a meaningful manner. It does not violate any of the modelling assumptions, yet it causes model instabilities and biases among the estimated effects of independent variables.

In this particular case, we will attempt to mitigate such biases via analysis of possible correlations among variables introduced in Section 3.2 based on our past game experience and mechanics knowledge.

Looking at the match descriptors, we expect no correlation since those variables only describe match information. However, we can not say the same for the independent variables. Logically, the game's outcome would affect gold and experience earned, e.g. a winning team would get more resources from killing more heroes and destroying a higher volume of buildings compared to the losing team. We are aware of this aspect, and variable *won* will be used to seek possible differences in the distribution of resources among players who won and lost and will be modelled individually for both cases. Another possible source of multicollinearity could arise from the relationship between *duration* and *consumables_purchased* since with lengthening duration, space for players to exchange money for such goods increases, and thus can create multicollinearity.

Possibly the most crucial challenge would appear while determining the correlation between gold and experience variables included. We believe that high levels of multicollinearity would occur if variables *GPM* and *XPM* will be present together with their respective distribution variables during the modelling phase. The argument is premised on the following logic: when an individual has higher *GPM* or *XPM*, sources of such commodity would be similarly upshifted because the spectrum of sources has to add up to present *total_gold* and *total_xp*. The presence of such two variables concluding total count likewise increases the multicollinearity across the model, and thus a closer look at the variable selection or a possible transformation has to be made further.

Also, we need to bear in mind that sole including both *GPM* and *XPM* can cause a measurable bias by itself. The reason is that gold increments go hand in hand with gains in experience - if a player obtains last hit, he will be given both resources. The same argument comes up in the case of gold and experience distribution since those two groups of variables will be tightly correlated - for example, in the case of *gold_roshan* and *xp_roshan* it is clear that the ratio would arise if and only if the team defeats Roshan and both metrics would be changing simoultaniously. The hypothesis should hold for all the similar variables in the pool.

4.3 Adressing such issues

As introduced at the beginning of this chapter, poor data quality was a prime problem. We wrote Python functions which did a great job handling *exceptions*[1] and during the mining process, we were able to insert *NaNs* where needed. After the dataset assembly, we deleted the data points missing crucial information, namely variables containing information regarding resources distribution, *APM* or *consumables_purchased*.

Section 4.2 describes the problem of multicollinearity in our dataset. To begin with, we realised some variables are causing perfect multicollinearity, breaking one of the elementary modelling assumptions not only for standard linear regression models. From Wooldridge (2009), this breach of assumption causes the matrix of the equation system representing our dataset not to have full rank, resulting in an impossible calculation of important linear algebra operations, i.e. inverse matrix calculation, valid only for matrices with full rank. The violating variables are *total_gold* and *total_xp*, since variables *GPM* and *XPM* are nontrivial linear combination of such predictors with *duration*. We had to omit such variables from the estimation to solve this problem.

Moreover, we decided to redefine our gold and experience distribution variables to be a number between zero and one, symbolising the share of total income obtained in a particular way. By doing so, a different source of perfect multicollinearity originating from relationship with *GPM* and *XPM* should cease to exist.

To support our recent decisions, we have generated a correlation heat map in Python. Figure 4.4 will be used primarily to identify strong correlations between predictors, from which we can make further decisions regarding our variable selection and have more information about the interrelationships amidst variables.

[1]Exception is a Python concept that occurs when something unexpected or undefined happens. For example, a division by zero or manipulating with non-existing information were valid exceptions in our mining step.

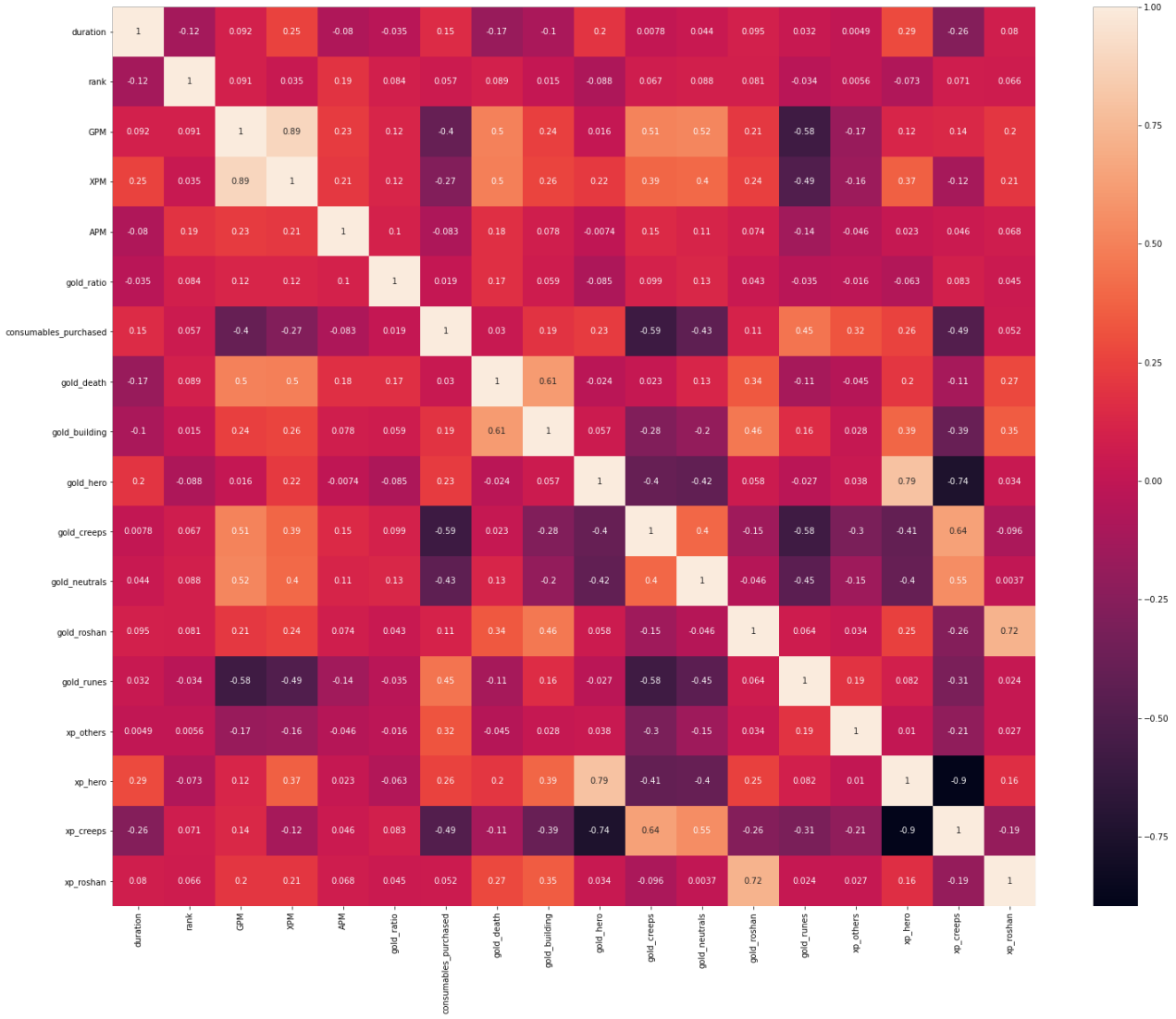

Figure 4.4: Correlation heatmap of the full variable pool

Returning to the multicollinearity-related notions mentioned above, it can be confirmed that their majority is correct. GPM and XPM variables are strongly interlinked with a correlation coefficient of 0.89. Further, it can be observed that the gold and experience distribution metrics seem to be strongly correlated with each other, resulting in the absolute value of correlation coefficient above 0.7 in numerous cases.

Based on the numbers, we believe including variables *XPM*, *xp_hero*, *xp_creeps* and *xp_roshan* generates multicollinearity, and this reduction in the number of regressors in our dataset is vindicated. We also decided to exclude variable *xp_others* since its effect on variable *rank* is rather small (correlation coefficient of 0.006), and the interpretation is much more difficult compared to other

distribution metrics. Besides, it was interesting to disclose the possible link between *consumables_purchased* and *duration* which was eventually refuted, and no thread of multicollinearity emerge from their joint presence.

It has to be noted that after heat map analysis of the regressand, the highest present value of correlation is circa 0.2, indicating possible low model performance and the validity of the null hypothesis.

We have finished the process of variable pool reduction, leaving us with the final 12 variables compared to the initial 17 predictors. It is likely that such predictor trimming had alleviated the possible biases and increased the credibility of outcomes.

Chapter 5

Estimation and data modelling

5.1 Algorithm selection and theory

In the modelling stage of the thesis, we will try to find models that best suit our dataset defined and crafted in Chapter 3 and Chapter 4. Since the dependent variable is categorical, standard linear regression is not applicable. We need to look towards a family of models based not on ordinary least squares, although on maximal likelihood. By this constraint, two models stand out - logistic and probabilistic regression.

Nevertheless, we need to modify such methods further to fit our data because fundamental logit and probit models work solely with binary variables. We need to extend those models to be polychotomous, implying operating with two or more dependent variable levels. Finally, we will obtain ordered logistic and probabilistic regressions, which are selected to explain the variation across our dataset.

To not have a monologic view on the problematicity, we have resolved to include another model operating on different principles. It is an AI algorithm called *K-Means Clustering* that uses possible information to seek resemblances and attempts to predict individual data partitions, making it a valuable tool for the analysis.

5.1.1 Logit and probit models

Logit and probit models are, as mentioned above, models used for datasets where the dependent variable is categorical and has more than two levels

which are *ordered*. Following detailed information about models' algorithms was sourced mainly from Hosein et al. (2015) and Wooldridge (2009), minor topics were discussed in study material from Stanford University by prof. Simon Jackman & Oscar Torres-Reyna from Princeton University. The estimation and interpretations differs from standard linear regression. If we look at the model equation, it does not differ drastically at the first sight:

$$y_i^* = \beta_0 + \beta_1 x_{1i} \cdots + \beta_k x_{ki} + u_i, \tag{5.1}$$

where β_0 is the intercept, $\beta_1 x_{1i} \cdots + \beta_k x_{ki}$ is a set of k independent variables and their respective effects, u_i is the error term and y_i^* is latent variable. This specific variable, meaning hidden in Latin, cannot be observed, and is used as a tool to classify the output into classes. The advantage of its unobservability is that we can assume it has specific shape. From Figure 3.1, we can see that the distribution of our dependent variable *rank* is similar to the normal distribution, so we can assume our latent variable's outcome has such shape and the center of the bell curve is the prediction. It has to be said that also the distribution of the error term, u_i, differs across logit and probit models. Using probit, error term would follow $N(0, 1)$. On the other hand, logit produces residuals which are logistically distributed. The classification itself can be described as follows (assuming we only have three classes, but the model can be applied to general problems where dependent variable has $m \in \mathbb{N}$ levels):

$$y_i = \begin{cases} 0 \;\; ; \;\; -\infty < y_i^* < \mu_1, \\ 1 \;\; ; \;\; \mu_1 < y_i^* < \mu_2, \\ 2 \;\; ; \;\; \mu_2 < y_i^* < \infty, \end{cases} \qquad i = 1 \ldots n \tag{5.2}$$

μ_1, μ_2 symbolize respective thresholds which divide the spectrum. In general case with m levels, $(m-1)$ thresholds will be present. Values of y_i are *Low*, *Mid* and *High* respectively in this analysis. Since the estimation output of logit and probit from Equation 5.2 is a probability, it makes a great sense to derive such relationships:

$$\begin{aligned} P(y_i = 1) &= P(\mu_1 < y_i^* < \mu_2) \\ &= P(\mu_1 < \beta_0 + \beta_1 x_{1i} \cdots + \beta_k x_{ki} + u_i < \mu_2) && (*) \\ &= P(\mu_1 - (\beta_1 x_{1i} \cdots + \beta_k x_{ki}) < u_i < \mu_2 - (\beta_1 x_{1i} \cdots + \beta_k x_{ki})) && (**) \\ &= \Phi(\mu_2 - (\beta_1 x_{1i} \cdots + \beta_k x_{ki})) - \Phi(\mu_1 - (\beta_1 x_{1i} \cdots + \beta_k x_{ki})) && (+) \end{aligned}$$

Other cases can be derived analogically:

$$P(y_i = 0) = \Phi(\mu_1 - (\beta_1 x_{1i} \cdots + \beta_k x_{ki})$$
$$P(y_i = 2) = 1 - \Phi(\mu_2 - (\beta_1 x_{1i} \cdots + \beta_k x_{ki}))$$

In Equation *, we used Equation 5.1 and subtracted the RHS with the exception of the error term in Equation **. The other two formulas for the dependent variable come from the fact that $y_i^* > -\infty$ always holds and probability of the highest group, *High rank*, is the complement of y_i^* being in the other two classes.

Across the two models, function Φ differs. In the logit case, function Φ is represented by Cumulative Distribution Function (CDF) for standard logistic random variable, a logistic function $\Lambda(x)$, defined as:

$$\Lambda(x) = \frac{\exp(x)}{1 + \exp(x)}. \tag{5.3}$$

On the other hand, Φ is represented by the CDF of standard normal distribution in the probit case. The CDF of $N(0, 1)$ follows formula:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(\frac{-t^2}{2})\, dt. \tag{5.4}$$

Both functions are defined, continuous with a range in interval (0,1), so the probability can be well-defined. If we would like to see the marginal effects of x_1, x_2 and x_3 (in the more general case also $x_4, \ldots, x_k$ if present in the model), we would take a partial derivative of particular probability with respect to the variable of choosing (for example take derivative of Equation + with respect to x_1 to obtain the marginal effect of x_1).

Because both models work differently from regular methods, so is the estimation. Linear regression uses Ordinary Least Squares (OLS) to obtain estimates, while the selected models are using MLE. Intuitively, MLE works in such a way that it iterates through the dataset with latent variable y_i^* and tries to fit the variable, in our case the bell shape, in a way that the likelihood of the data points is maximized, meaning most of the values are located near the tip of the bell, or the mean value, as shown in Figure 5.1. Mathematically, this process

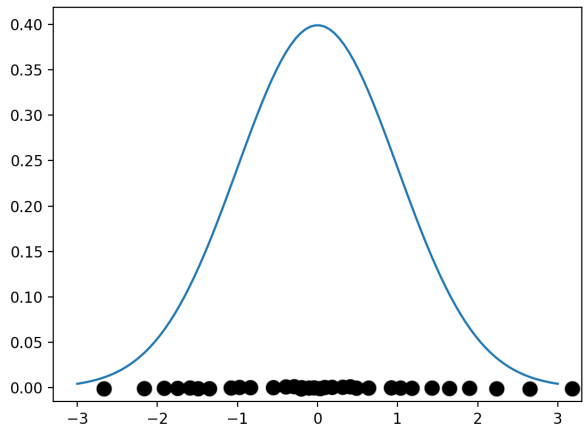

Figure 5.1: Bell curve with high density of observations around the mean, an example of MLE fit.

is done via maximization of log-likelihood function, defined as:

$$\mathcal{L}(\beta) = \sum_{j=1}^{J}\sum_{i=1}^{n} y_i \ln(\Phi(\mu_j - x_i\beta) - \Phi(\mu_{j-1} - x_i\beta)) + \\ +(1-y_i)\ln(\Phi(\mu_j - x_i\beta) - \Phi(\mu_{j-1} - x_i\beta)),$$

where function Φ differs for logit and probit as mentioned above.

In order to make statements about the performance of the models, a relevant metric has to be introduced. Due to the nature of both models' outcomes, it is not possible to use the standard R^2 formula. Since there is no best-performing model evaluation, we chose two Pseudo-R^2 metrics to express the amount of variation our models explain: McFadden's R^2 and Cox & Snell's R^2.

Intending to compare such metrics, we will state differences across their formulas. Firstly, from McFadden & other (1973), McFadden's R^2 is defined as follows:

$$\rho^2 = 1 - \frac{\mathcal{L}(M_{full})}{\mathcal{L}(M_{null})}, \tag{5.5}$$

where $\mathcal{L}(M_{full})$ and $\mathcal{L}(M_{null})$ represent log-likelihood of the full and null model, respectively. Secondly, Cox & Snell's R^2 is measured as:

$$R^2 = 1 - \left(\frac{\mathcal{L}(M_{full})}{\mathcal{L}(M_{null})}\right)^{\frac{2}{n}}, \tag{5.6}$$

obtained from Cox & Snell (1989). Again, $\mathcal{L}(M_{full})$ and $\mathcal{L}(M_{null})$ represent log-likelihoods arising from full and null models, n is the number of observations. The main difference is the n^{th} root of the squared likelihood ratio in formula for Cox & Snell's R^2. As the number of observations increases, it drives up the degree of the root, and thus makes models with large dataset better-performing, on contrary to the McFadden's, which only takes in consideration the likelihood ratio.

5.1.2 K-Means Clustering

K-Means Clustering is a part of Unsupervised Machine Learning family of algorithms, meaning that no other action apart from specifying required parameters is needed; a computer does conclusions independently. On the contrary, Supervised Learning algorithms such as K-Nearest Neighbors, Decision Trees or ANNs require human supervision of the output and the process of tuning the hyperparameters is done by the human. Unsupervised Learning algorithms are most commonly used to find patterns across the data. An example of a related algorithm is, for example, Hierarchical clustering. We believe that it would be of great interest to examine if a computer can structure our DotA 2 dataset into somewhat similar groups to the rank division within if we set our $k = 3$.

The algorithm works as follows: initially, it randomly spreads k cluster centres across the dataset. Next, it assigns respective data points to the nearest cluster, obtained by Euclidean distance. The following formula computes this distance:

$$\|\mathbf{p}, \mathbf{q}\| = d\left(p, q\right) = \sqrt{\sum_{i=1}^{n} \left(q_i - p_i\right)^2}, \tag{5.7}$$

where p and q are desired points, and n symbolises the points' dimension. After the initial segmentation, centroids are re-calculated to be situated in the actual centre of the cluster and re-calculates the points belonging to the new clusters. This process is repeated until convergence or until the maximum possible iterations are achieved and presents the results by a vector of numbers corresponding to individual observations.

We will assess the results produced by k-means algorithm by simply calculating the accuracy of the model, defined as:

$$\text{accuracy} = \frac{\text{sum of correctly guessed clusters}}{\text{total number of observations}}. \tag{5.8}$$

5.2 Estimation

Since the gold and experience gains depend heavily on the game's outcome and hero position within the team, we decided to split our dataset into four subsets - carry players who won, carry players who lost, support players who won and support players who lost. Due to this segmentation, we believe that the effects on rank would be much more evident, and the estimation will be of greater research value. When interpreting the individual datasets, we will focus on finding patterns across the summaries and seeking possible differences among the groups.

Applying the variable reduction from Section 4.3, we have finally obtained the model structure sufficient enough to be marked final. We have tried both ordered logit and probit models with almost identical results. We finally decided to use the logit model thanks to its Φ function simplicity and calculation.

The estimation coefficients confirmed the hypotheses we set up in Section 3.2. Moreover, all the effects were strongly statistically significant, indicating the possible answer to the thesis question.

As we can see from the Table 5.1, the ordered logit model produces not only the estimates for individual independent variables, yet it has a second part denoting intercept values. Those intercepts symbolize the μ_1 and μ_2 threshold values for the latent variable y_i^* to define the intervals for rank clusters and assignment of final y_i value, previously defined in Equation 5.2. Extremely low p values and Standard Errors also require further explanation. Since DotA 2 games included in our dataset were filtered to avoid extreme values, our dataset would have dispersion within borders defined by a single DotA 2 game. Concretely from Figure 4.2, assumptions about the rough normality of the regressors can be made. Consequently, independent variable values also lie inside constrained boundaries, and since every game happen independently of

Table 5.1: Ordered logit summary of *carry_won*

Dataset ***carry_won***		27 542 observations	
Response variable	**Coefficent**	**Standard Error**	**p value**
duration	-0.001328	4.109e-05	5.523795e-229
APM	0.003176	1.250e-04	2.319735e-62
GPM	0.001880	1.128e-04	1.570693e-142
gold_ratio	0.979490	2.270e-04	0.00
consumables_purchased	0.033901	1.952e-03	1.400132e-67
gold_death	-5.039122	2.272e-05	0.00
gold_building	-3.194761	3.400e-05	0.00
gold_hero	-2.116227	4.811e-06	0.00
gold_creeps	2.462607	1.004e-04	0.00
gold_neutrals	1.013808	7.109e-06	0.00
gold_roshan	27.518800	6.731e-05	0.00
gold_runes	-7.941796	1.615e-05	0.00
Intercept			
0|1	-1.1605	0.0002	0.00
1|2	0.7657	0.0181	0.00
McFadden's R^2		**5.4%**	
Cox & Snell's R^2		**10.1%**	

Note: This table represents the whole set of summaries in main text. Remaining three tables with results from other datasets can be viewed in Appendix A.

other games, no major discrepancies occur, resulting in minuscule p values and standard errors.

The portrayed logit results in Table 5.1 signalise possible evidence to conclude that defined metrics are sufficient predictors of the rank clusters. Nonetheless, this changes as we calculate model assessment metrics. We extracted McFadden's R^2 together with Cox & Snell's R^2, resulting in the average values of 6.2% and 11.7% from all the four model summaries, respectively. This finding changes the output perception because only a little information has been explained across the data. To make final statements, we will study the phenomenon further with the cluster analysis using Unsupervised Machine Learning algorithm, *K-Means Clustering*.

Similarly to the logit estimation, the K-Means procedure has been applied to each dataset to eliminate unwanted trends resulting from previously mentioned boolean variables *won* and *is_carry*, shown in Table 5.2. Nevertheless, direct algorithm output was insufficient to make conclusions, further manipulation with the clusters had to be made. Firstly, individual clusters and rank groups comparison has been made to match the clusters with the most similar rank segment. The clustering accuracy has been computed after matching clusters with the most alike *rank* group via exploration of independent variables' mean values. As in the logit modelling, AI's performance was not particularly solid. We have obtained 37% accuracy from *carr_won* subset, 41% from *carr_lost*, 37% from *supp_won* and 38% accuracy in *supp_lost*. Consequently, it leads us to a mean accuracy of 38.25% and weighted[1] mean accuracy of 38.4%, only a scant improvement compared a random classifier.

Table 5.2: Cluster analysis results

Dataset	***carry_won***	***carry_lost***	***support_won***	***support_lost***
Observations	27 542	27 526	18 404	18 388
Accuracy	37%	41%	37%	38%
$\overline{\textbf{Accuracy}}$			**38.25%**	
$\overline{\textbf{Accuracy}}_{weighted}$			**38.4%**	

[1]With respect to the number of observations.

Chapter 6

Results

To conclude whether the researched effect exists, it would be desirable to begin with an explanation of how to handle the output of polychotomous logistic regression. As previously discussed in Subsection 5.1.1, ordered logit does not present the direct result, yet it tells us how it affects the value of the dependent variable, y_i^*. The process is as follows: It receives an i-th row, multiplies all the variable coefficients and respective actual values, sums and presents the final value of y_i^*. Based on y_i^*, it calculates the odds that the dependent variable lies within each cluster based on threshold values of μ_1 and μ_2 included in the summary in the following manner:

$$P(y_i = 0|\mathbf{x}) = \frac{1}{1 + \exp(-(\mu_1 - y_i^*))}. \tag{6.1}$$

To better understand the equation, it can be interpreted as the probability of a certain data point being labelled as *Low* rank, given $\mathbf{x}$, all the information regarding the data point. Same logic applies to probabilities of lying in other *rank* groups. Further, y_i^* represents the Equation 5.1. This yields a well-defined probability between 0 and 1, representing the likelihood that this particular observation is marked as *Low* rank. However, we will interpret only signs of the estimations and their implications for a real match.

Looking at the summaries, we acknowledge a negative impact of *duration* likely caused by the fact that better players are more accurate at identifying the point of the game when the gap between teams is fatal and thus focus all the attention to close the game as soon as possible. This argument is further supported by Table 6.1, from which can be deduced that the difference between

High and *Low* games is on average more than a minute, being marginal between *Mid* and *Low* ranks.

Table 6.1: Mean variable values for *duration*

Duration	**Mean value**
High	1740.71
Mid	1798.48
Low	1813.12

Following with *GPM* and *APM*, we can observe a positive correlation with dependent variable. Our reasoning behind such sighting is that the ability of more skilled players to extract resources more rapidly together with better time management can result in a *GPM* improvement. We can likewise defend variable *APM* as the relationship roots in faster reactions, better focus and generally more presence in the game for experienced players. Our conclusion regarding *APM* is also supported outside of this thesis by Castaneda et al. (2016), as it describes the increasing importance of shortcuts and their usage among better players.

Table 6.2: Mean variable values for *GPM* and *APM*

GPM	**Mean value**	***APM***	**Mean value**
High	246.12	*High*	200.71
Mid	243.65	*Mid*	178.27
Low	240.23	*Low*	157.86

Shifting our focus on spending habits, effects of *consumables_purchased* and *gold_ratio* will be explained. As was predicted earlier in Section 3.2 and supported by Table 6.3, players with expertise should be more reactive to the current situation, and thus will focus on purchasing more consumables of smaller value, yet having a greater immediate effect which can allow them to say alive longer compared to beginners, who will most likely see the greatest utility gain from expensive items, imposing a much greater risk to themselves. Our predictions connected to the behaviour of *gold_ratio* have also come to fruition, as

the players tend to be more effective in spending their resources with increasing insight. En plus, it was correct to assume an ideal value of 0.9, as players tend to safe gold for buybacks, earlier mentioned in Section 3.2.

Table 6.3: Mean values for *consumables_purchased* and *gold_ratio*

consumables_purchased	**Mean value**	***gold_ratio***	**Mean value**
High	25.19	*High*	0.88
Mid	22.55	*Mid*	0.85
Low	19.54	*Low*	0.83

A great deal of interesting information can be gleaned from individual summaries and Table 6.4 containing gold distribution metrics and their variation across carry and supports players, underlying their in-game roles.

Table 6.4: Mean values for gold distribution variables

gold_death	**Mean value**	***gold_creeps***	**Mean value**
High	-0.1002	*High*	0.1494
Mid	-0.1076	*Mid*	0.1483
Low	-0.1152	*Low*	0.1522
gold_neutrals	**Mean value**	***gold_hero***	**Mean value**
High	0.03807	*High*	0.1998
Mid	0.03700	*Mid*	0.2088
Low	0.03454	*Low*	0.2103
gold_building	**Mean value**	***gold_runes***	**Mean value**
High	0.026	*High*	0.073
Mid	0.023	*Mid*	0.071
Low	0.021	*Low*	0.068

Firstly, we can see a logical positive trend indicating a lower number of deaths with increasing rank for most subsets with the exemption of supports who lost. We believe this is caused by the fact that *High* rank supports can realise the game is not unfolding according to the desired scenario and thus

try to create space for the carry players to catch up, sacrificing themselves if needed.

Secondly, we can observe a different patterns in *gold_creeps* and *gold_neutrals* between carry and support datasets. While high-skilled carry players average higher gold proportions from those sources, support players tend to decrease the share with higher ranks. It is due to different obligations those roles have. The objective of support is to ensure that carry players have all the space to earn gold and experience. In addition to the overall negative relationship of *gold_hero*, it points out that top players are being risk-averse thanks to their focus on obtaining gold from neutrals and creeps instead of committing energy to kill heroes, which poses a significant risk of dying. This information is not extractable form Table 6.4 since the effect is combined with support data, yet it can be seen in the summaries. To confirm this, we have created Table 6.5 depicting the phenomenon in more detail.

Table 6.5: Carry and support differences in *gold_creeps* and *gold_neutrals*

Carry *gold_creeps*	**Mean value**	**Support *gold_creeps***	**Mean value**
High	0.3312	*High*	0.1341
Mid	0.3164	*Mid*	0.1368
Low	0.2978	*Low*	0.1426
Carry *gold_neutrals*	**Mean value**	**Support *gold_neutrals***	**Mean value**
High	0.1351	*High*	0.0352
Mid	0.1196	*Mid*	0.0359
Low	0.1080	*Low*	0.0348

Thirdly and finally, the effect orientation of last two metrics *gold_building* and *gold_runes* on *rank* is dependent on the winning condition. We believe this can be translated as the volume of gold obtained from buildings and runes is approximately the same, with the main difference in total gold obtained is trivially greater if a player wins a game, causing the negative slope in the case of lost games.

Yet possibly the greatest impact on *rank* is generated from *gold_roshan*. From Table 6.6, the ratio is 56% higher for *High* rank players compared to

Low and difference of 34% from *Mid* to *High*. Such a lacuna can arise from beginners' inability to identify a moment in which they are sufficiently powerful to defeat a creature as mighty, or merely the fact that they get carried out by the heat of the game, forgetting the Roshan's existence.

Table 6.6: Mean variable values for *gold_roshan*

gold_roshan	**Mean value**
High	0.00169
Mid	0.00113
Low	0.00074

Chapter 7

Conclusion

The objective of this thesis was to examine if it is possible to predict the rank of an individual DotA 2 player solely based on data regarding his or her consumer behaviour. We have implemented two methods, ordered logistic regression and cluster analysis via an AI algorithm K-Means Clustering.

Mentioned tools have not proven there is a strong influence of consumer behaviour on the rank of individuals playing DotA 2. We could observe that logit estimation explained roughly eleven per cent of the dataset variation, and cluster analysis has resulted in a mean accuracy of 38.25% and weighted mean accuracy of 38.4%. The null hypothesis is also favoured by the heat map, indicating the strongest correlation with the dependent variable of roughly 0.2. We believe that the utils used to examine the dataset were chosen correctly, and also the data was not of low quality - majority of the included variables have well-behaved dispersion, resembling almost normal distribution.

Evidence from the previous chapter presents a relatively precise overview of what this research has discovered. We can conclude that consumer behaviour changes with rank, yet the effect is not sufficient to determine the rank solely by itself, as seen from the algorithm assessments and evaluation of the results. Especially the K-Means Clustering depicts how effectively patterns are found across the data and how strong the evidence about the rank is when we conceal it.

Bibliography

Akhmedov, K. & Phan, A. H. (2021). Machine learning models for dota 2 outcomes prediction.

Beskyd, F. (2018). Prediction of DotA 2 game result.

Castaneda, L., Sidhu, M., Azose, J., & Swanson, T. (2016). Game play differences by expertise level in DotA 22, a complex multiplayer video game. *International Journal of Gaming and Computer-Mediated Simulations*, *8*, 1–24.

Cox, D. & Snell, E. (1989). *Analysis of Binary Data, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Donaldson, S. (2017). Mechanics and metagame: Exploring binary expertise in league of legends. *Games and Culture*, *12*(5), 426–444.

Drachen, A., Yancey, M., Maguire, J., Chu, D., Wang, I. Y., Mahlmann, T., Schubert, M., & Klabajan, D. (2014). Skill-based differences in spatio-temporal team behaviour in defence of the ancients 2, 1–8.

Hosein, M., Sasan, T., & Dogani, A. (2015). Application of ordered logit model in investigating the factors affecting peopleâ€™s income (a case study in tehran city). *International Journal of Academic Research in Economics and Management Science)*, *5*(3), 166–178.

Katona, A., Spick, R., Hodge, V. J., Demediuk, S., Block, F., Drachen, A., & Walker, J. A. (2019). Time to die: Death prediction in dota 2 using deep learning, 1–8.

Kočur, J. (2018). Umělý hráč pro dotu 2.

Kwak, H., Blackburn, J., & Han, S. (2015). Exploring cyberbullying and other toxic behavior in team competition online games. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

McFadden, D. & other (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in econometrics*, 105–142.

Mora-Cantallops, M. & Ángel Sicilia, M. (2018). MOBA games: A literature review. *Entertainment Computing*, *26*, 128–138.

OpenAI et al., . (2019). DotA 2 with large scale deep reinforcement learning.

Rümmele, N., Neidhardt, J., Calatrava Moreno, M. d. C., Grad-Gyenge, L., & Werthner, H. (2013). On successful team formation: Statistical analysis of a multiplayer online game, 55–62.

Simon, F. (2014). From generative to conventional play: MOBA and league of legends.

Tseng, F.-C. (2011). Segmenting online gamers by motivation. *Expert Systems with Applications*, *38*(6), 7693–7697.

Winn, C. (2015). The well-played MOBA: How DotA 2 and league of legends use dramatic dynamics.

Wooldridge, J. M. (2009). *Introductory Econometrics: A Modern Approach*. ISE - International Student Edition. South-Western.

Xia, B., Huiwen, W., & Ronggang, Z. (2019). What contributes to success in moba games? an empirical study of defense of the ancients 2. *Games and Culture*, *14*(5), 498–522.

Yang, P., Harrison, B. E., & Roberts, D. L. (2014). Identifying patterns in combat that are predictive of success in MOBA games.

Appendix A

Additional ordered logit summary tables

Table A.1: Ordered logit summary of *carry_lost*

Dataset ***carry_lost***		27 526 observations	
Response variable	**Coefficient**	**Standard Error**	**p value**
duration	-0.001989	4.141e-05	0.00
APM	0.004011	1.592e-04	4.4971e-140
GPM	0.003450	1.395e-04	5.1610e-135
gold_ratio	0.188126	2.472e-04	0.00
consumables_purchased	0.29685	1.839e-03	1.256e-58
gold_death	-1.168802	3.611e-05	0.00
gold_building	11.475116	8.525e-06	0.00
gold_hero	-1.197861	3.476e-05	0.00
gold_creeps	5.380855	1.510e-04	0.00
gold_neutrals	3.808568	3:271e-05	0.00
gold_roshan	25.230599	8.703e-07	0.00
gold_runes	9.323111	2.112e-05	0.00
Intercept			
0|1	0.5596	0.0003	0.00
1|2	2.5646	0.0186	0.00
McFadden's R^2		**8.2%**	
Cox & Snell's R^2		**15.1%**	

Table A.2: Ordered logit summary of *support_won*

Dataset ***support_won***		18 404 observations	
Response variable	**Coefficent**	**Standard Error**	**p value**
duration	-0.001676	6.716e-05	1.762e-177
APM	0.003385	1.675e-04	1.316e-27
GPM	0.003206	2.944e-04	7.792e-91
gold_ratio	0.280574	2.468e-04	0.00
consumables_purchased	0.024574	1.222e-03	1.238e-90
gold_death	-2.261877	3.027e-05	0.00
gold_building	-1.995705	6.353e-05	0.00
gold_hero	-0.579391	3.771e-05	0.00
gold_creeps	-1.838048	3.628e-05	0.00
gold_neutrals	-0.445654	6.350e-06	0.00
gold_roshan	29.355530	1.317e-07	0.00
gold_runes	-3.891067	3.558e-05	0.00
Intercept			
0|1	-2.6582	0.0003	0.00
1|2	-0.7879	0.0216	2.588e-292
McFadden's R^2		**4.7%**	
Cox & Snell's R^2		**9%**	

Table A.3: Ordered logit summary of *support_lost*

Dataset ***support_lost***		18 388 observations	
Response variable	**Coefficent**	**Standard Error**	**p value**
duration	-0.00203	6.129e-05	2.475e-234
APM	0.003919	1.859e-04	1.226e-127
GPM	0.005375	3.700e-04	8.066e-48
gold_ratio	0.601117	3.285e-04	0.00
consumables_purchased	0.031699	1.320e-03	2.079e-127
gold_death	2.02168	2.914e-05	0.00
gold_building	11.297789	1.858e-05	0.00
gold_hero	-0.538964	3.896e-05	0.00
gold_creeps	0.240943	2.882e-06	0.00
gold_neutrals	1.133526	6.802e-06	0.00
gold_roshan	20.390797	1.308e-06	0.00
gold_runes	6.386785	5.217e-05	0.00
Intercept			
0|1	-1.6241	0.0004	0.00
1|2	0.3627	0.0222	5.090e-60
McFadden's R^2		**6.6%**	
Cox & Snell's R^2		**12.5%**	

Appendix B

Additional sources

In this part of the Appendix, alternative sources to academic research used in text will be listed.

1) *IOC makes landmark move into virtual sports by announcing first-ever Olympic Virtual Series.* International Olympics Comittee. Available from `https://bit.ly/3eTMhrW`.

2) *Valve Corporation.* Valve software, January 2013. Available from `http://www.valvesoftware.com/`.

3) *The International.* Liquipedia DotA 2 Wiki. Available from `https://liquipedia.net/dota2/The_International`.

4) *Largest Overall Prize Pools in Esports.* Esport Earnings. Available from `https://www.esportsearnings.com/tournaments`.

5) *Summary of Active DotA 2 players.* Steam Charts. Available from `https://steamcharts.com/app/570`.

6) *Models for Ordered Outcomes.* Simon Jackman, Political Science 200C course, Stanford University study material. Available from `https://web.stanford.edu/class/polisci203/ordered.pdf`.

7) *Getting Started in Logit and Ordered Logit Regression.* Oscar Torres-

Reyna, Princeton University. Available from http://pioneer.netserv.chula.ac.th/~ppongsa/2900600/LMRM06.pdf.