Master thesis

Data Science and Marketing Analysis

Analysing and predicting peer behaviour in a gaming environment: a big data analysis of moderator effects

ERASMUS UNIVERSITY ROTTERDAM,
Erasmus School of Economics

In literature, there exists a gap between peer effects and gaming behaviour. This research aims to close that gap and, in doing so, determine what factors affect gaming behaviour. It became clear that gaming behaviour of a gamer is first, and foremost, affected by gaming behaviour of its peers. However, this relationship is severely moderated by other characteristics. These characteristics are separated into relationship characteristics and game characteristics. The linear regression model demonstrated that number of peers and similarity between peers are important characteristics of the relationship, whereas price, DLC and online capabilities seem to be important factors of the characteristics of the game. In addition to that, the comparison between a binomial regression and a Random Forest demonstrated that the latter is most capable of predicting peer effect in a gaming environment.

Name student: Martijn Groenendijk
Student ID number: 387418

Supervisor: Dr. R. Karpienko
Second assessor: Prof.dr.ir. R. Dekker

Date final version: February 5th, 2021

I. Preface

The work contained in this thesis has not been previously submitted for a degree or diploma at any other higher education institution. To the best of the author's knowledge and belief, the thesis contains no material previously published or written by another person except where due references are made.

Data collection of this thesis is performed in two different manners. The data that is gathered from a previous study, is owned by the writers of that study and is acquired for the purpose of this thesis only. Reproduction or distribution of the data is not permitted, only with the authorisation of the authors of the study. Also, the data collected through the APIs of Steam is for the purpose of this thesis only. The information of the Steam users will be used by the staff of ESE for examination only.

II. Acknowledgement

This thesis marks the end of my academic career at the Erasmus School of Economics. It has been an incredible journey that started six years ago. Back then, I did not have an obvious career path in mind. Because of the broad future possibilities it offers, I enrolled in the Bachelor programme Economics & Business Economics. During my bachelor's I learned a lot about my interests and, after four years, I proceeded to the next step in my academic career. Now that I am about to graduate from the Master Data Science and Marketing Analysis, I can say with confidence that my professional career path has become clear.

Writing this thesis has played a valuable part in my academic development. First of all, it represents the ultimate task of connecting the dots because it trigged me to use all important skills that I have learned throughout my bachelors and masters. During the past six months, I was challenged to broaden my analysing, coding, researching and writing skills. In addition to that, this thesis has enabled me to combine my passion for economic concepts with big data analysis. At the same time, it offered me the possibility to apply both topics to a technological environment that has my interest.

It is important to mention that this thesis would not have been such a valuable process without the help of several important people. Therefore, I would like to take this opportunity to thank these people. First of all, I would like to thank my University supervisor Radek Karpienko for the valuable insights he has given me. He provided me with constructive feedback on important parts of my thesis, as well as detailed examples of how to tackle my ever-expanding and challenging process of data collection.

Besides that, I would like to thank my family and friends. Mom, thank you for the unconditional support, love and confidence. You provided me with the strong foundation that I needed during my academic career. Michelle, thank you for the goofy times and, especially, for your high-value input during my bachelors and masters. It has enabled me to reach an even higher potential. Last but not least, I would like to thank my friends for enjoying the good times and celebrating our accomplishments.

I am looking back at this process with a great sense of proudness and I am looking forward to the next chapter of my career.

Martijn Groenendijk, Rotterdam, February 2021

III. Executive summary

In literature, there exists a gap between peer effects and gaming behaviour. This research aims to close that gap and, in doing so, determine what factors affect gaming behaviour. It became clear that gaming behaviour of a gamer is first, and foremost, affected by gaming behaviour of its peers. However, this relationship is severely moderated by other characteristics. These characteristics are separated into relationship characteristics and game characteristics.

Literature suggests that relationship characteristics consist of number of peers of a gamer and similarity between gamers. Furthermore, the game characteristics consist of the price of a game, the amount of downloadable content, the number of achievements and the recommendations of a game. In addition to that, several genres and online capabilities are identified. Because of the scientific nature of this thesis, it follows a predetermined structure.

First, the history of the gaming industry is discovered, followed by the current state of this industry. In doing so, it became clear that the exact factors, that influence gaming behaviour, are unknown. This forms the basis for the research question and four additional sub-questions, which focus on moderator effects that affect gaming behaviour, and to what extent peer effects can be predicted by a binomial regression or Random Forest.

Secondly, all literature that is relevant for the relationship and game characteristics is provided. Based on this literature, several hypotheses are given to answer the sub-questions appropriately. After that, the process of data collection is described. The data in this thesis is collected from a previously conducted research and through the APIs of Steam's gaming environment. Altogether, the data collection resulted in a data set with 22 variables of 260,389 different players. This data is used by a linear regression, binomial regression and Random Forest to test the hypotheses.

The linear regression model demonstrated that number of peers and similarity between peers are important characteristics of the relationship, whereas price, DLC and online capabilities seem to be important factors of the characteristics of the game. In addition to that, the comparison between a binomial regression and a Random Forest demonstrated that the latter is most capable of predicting peer effect in a gaming environment. This thesis concludes with an explicit answer to the research question, the limitations of this research and the directions for further research.

IV. Table of contents

V. List of figures and tables

1. Introduction

1.1 Historical background of the gaming industry

The gaming industry has seen rapid development over the past few decades. Where it used to be a small - and maybe irrelevant - industry that provided gaming halls with arcade machines, today it is one of the most innovative technology sectors. The first commercial release of a video game dates back to the 1950s. The release of video games, such as Tennis for Two, was regarded as just 'another' peculiar technological invention. During the 1950s and 1960s, the gaming industry kept this stigma. In the late 1960s and early 1970s, companies as Sega and Atari – which are known as major gaming companies, especially during the development of the gaming industry – released their first electronic video game (History.com, 2019).

Atari was the first company that showed the real possibilities of gaming machines with famous games, such as Pong and Asteroids. This led to the initial development of gaming communities. Besides home video game consoles, Atari also founded the arcade gaming industry (Atari, n.d.). These gaming systems were found in bars, shopping malls and game halls. At that point, other companies started realizing that gaming could be 'the next big thing'. During the mid-1970s and mid-1980s, more than a dozen new gaming companies emerged that all wanted a piece of the pie. This triggered the start of the rapidly evolving gaming industry (Chikhani, 2015).

This evolution encountered a major boost when restaurants installed arcade gaming machines in the late 1970s. Due to the way these games were built, the user experienced a sense of competition. Gamers tried to outrank each other by breaking the high score. In this manner, the gamers' initials would be on top of the list. This development is seen as the birth of multiplayer games. At that moment, multiplayer games were limited to playing against each other on one screen. The possibility to play on separate screens was first released in the early 1980s. However, due to the high costs of the computers and network that were required for such games, access to this multiplayer experience was limited to major companies and universities. (Chikhani, 2015).

In the 1980s and 1990s, the development of the gaming industry progressed steadily. Game consoles saw inventions such as an interchangeable cartridge, instead of a preloaded set of games, that offered the possibility for users to build their video game library. As a result, software developers could join the industry and develop games without selling their own gaming device. Due to the rapid increase in games, the market became saturated and the quality of games began to fall. This led to a crash of the market in the early 1980s. Besides that, video consoles were confronted with competition from the personal computer. Devices such as the Commodore 64 and Apple II

were publicly available, and households were able to justify the investment because it allows for much more than video gaming (Smithsonian, n.d.).

These Personal Computers (PCs) had – besides a more justifiable purchase decision – more powerful processors than the previous generation game consoles. Video games became less linear and more complex due to 3D-gaming possibilities for example. Additionally, PCs were more 'open-source' than game consoles, meaning that PC users could create their own games as well as designing possibilities for multiplayer gaming. In the late 1980s, PC owners, but also console owners, started to connect devices so they could play games with each other on two separate devices. The release of the 'LAN Party' in 1993 is a true turning point in the multiplayer experience of gaming. This type of multiplayer gaming offered gamers to connect up to 16 devices in a Local Area Network (LAN) (Medium.com, 2017). From the late 1990s, the possibilities in terms of online capabilities grew rapidly. With developments as World Wide Web service and inexpensive ethernet cards, online gaming was well on its way to became as we know it today.

In the early 2000s, computers and consoles became even more powerful, and internet connectivity was built into game consoles. In the decades that follow, the costs of technological development decreased significantly, making it easier for new game devices to outclass the previous generation. In addition, it had become easier for people to gain access to the internet. In 2000, 413 million people had access to the internet and in 2020 this number has increased to 4.66 billion people (Clement, 2020). Reports show that in 2020, at least 1.9 billion people with internet access, play games with online capabilities (The Entertainment Software Association (ESA), 2020).

1.2 The current state of the gaming industry

The powerful rise of the gaming industry has led to several new developments over the last decade. Online capabilities have become more integrated and woven into the fabric of games. This has changed how gamers buy games, update games and interact with other gamers. Gaming devices such as the PlayStation, Xbox and PC, each offer their own platform that facilitates a game store, as well as the online capabilities of a game. Popular examples of such platforms are Sony's PlayStation Network (PSN), Microsoft's Xbox Live and Valve's Steam. These gaming platforms have shifted the focus of gaming towards a more social activity. Gaming used to be stigmatized as an anti-social activity. However, in more recent years, gaming is regarded as an activity that supports social interaction (Eklund, 2012). According to the ESA (2018); 55 per cent of gamers say that gaming helps to connect with friends and 50 per cent of gamers feels that it helps the family to spend more time together.

Today, most game developers offer games with online gameplay that surpass the offline gameplay. In that sense, a good multiplayer experience has become the main objective for a developer. As a result, the majority of the global revenue of the gaming industry is generated through multiplayer components. According to Statista (n.d.a) over one-third of the population plays video games, generating a global revenue of over 150 billion US dollars. This is a growth of 5.2 per cent compared to last year. In 2020 the mobile gaming segment generated 77.2 billion US dollars, followed by 45.2 billion US dollars and 36.9 billion US dollars for the console gaming segment and PC-gaming segment respectively (Statista, n.d.b). Some of the major players on the market are Tencent, Sony and Apple, that have combined revenue close to 11 billion US dollars in 2019 (NewZoo, n.d.a). However, for this research, only the PC-gaming segment is taken into consideration, which will rule out some publishers and developers as contributors. This is because several companies only develop and publish on one device specifically.

1.3 Problem definition

As described in the previous sections, the gaming industry has grown rapidly over the past decades. Besides that, it is demonstrated that it has become part of our daily life and helps with social interaction. Making friends is an important skill for people of all ages, but especially for young children. According to Berndt (2002), high-quality friendships have a direct effect on positive characteristics, such as prosocial behaviour and is a good predictor of happiness (Argyle, 2001). Therefore, most people have an intrinsic motivation for seeking and maintaining friendships. Additionally, some friendships induce behaviour where one peer influences the attitude and actions of another peer. These effects that cause people to change their behaviour because of one's peer group, are called peer effects. Therefore, some individuals might invest more time into gaming than others because of these peer effects.

Besides that, game designers have found various methods to make games more appealing throughout the years. The research of Zagal, Björk, & Lewis (2013) argues that – even though game designers are usually regarded as "advocates for players" – some games are designed with dark patterns. These patterns are "negative gaming experiences, built into the game on purpose, with or without the gamers' consent and against their best interest". The experiences include social capital-based dark patterns, temporal dark patterns and monetary dark patterns.

Social capital-based dark patterns consist of social pyramid schemes, as well as a game impersonating a player. In social pyramid schemes, players are rewarded based on the in-game performance of their friends. In turn, these friends are rewarded for the in-game performance of

their friends. This process can continue indefinitely, resulting in extreme high rewards for those on top of the pyramid and it magnifies peer effects. Further, when a game impersonates a player, it sends a message to the friends of a given player. This message aims to nudge other players into investing time or money in that particular game because their friends also did. In this manner, the game creates an artificial peer effect on that player. This implicates that games, designed in such manner, take advantage of players and make them more susceptible to peer effects

Temporal dark patterns focus on "cheating players out of their time". In doing so, players have to participate in repetitive gaming (i.e., "grinding") or have to play the game according to a predetermined schedule. Furthermore, a game designer can implement monetary dark patterns into its game. These patterns can include "pay-to-skip' strategies that nudge players to invest money so they can make progress. Also, pre-delivered content is a monetary dark pattern. This pattern encourages gamers to pay an additional fee on top of the initial price to unlock certain parts of the game. Even though this might not seem like a dark pattern, it is a malignant manner to ask more money for content that was already on the disk in the first place. Besides that, players can be encouraged to invest more money through "monetized rivalries". This is also known as the "pay-to-win" strategy, as it exploits the competitiveness of players.

In sum, a game, as well as the relationship between two players, can be designed in such manner that it encourages gamers to invest more time and money. These characteristics can either have a direct effect on a gamer or indirect on the friends of a gamer through peer effects. For game designers, it is highly valuable to analyse the implications of these characteristics and to derive meaningful insights from them.

1.4 Steam

The focal point of this thesis is the gaming environment of Steam. This company focusses on online distribution of PC-games. Because this gaming environment is easily accessible for all sizes of game developers - and offers all sorts of tools to improve distribution for publishers - Steam counts over 30,000 PC-games, ranging from AAA games to indie games. Besides distributing games, Steam also provides its users with the possibility to participate in a gaming community. A community allows for meeting fellow gamers, joining groups, creating clans and facilitating in-game chats. As of 2019, the Steam community counts over 100 million registered users. Additionally, Steam offers gaming hardware, such as VR equipment and mobile gaming devices (Steam, n.d.b). Steam is established in 2003 and owned by the Valve Corporation. Valve mainly develops and publishes games and, in turn, uses Steam to distribute their games (Valve Corporation, n.d.a).

1.5 Research question

The main question that arises from the existing literature is: do certain factors alter gaming behaviour in a game environment? These factors can either be connected to the characteristics of a relationship between two gamers, or the characteristics of the game. Either way, game designers can benefit from these insights and built certain characteristics into a game. In turn, gamers might invest more time and money into a specific game. This results in game designers that are better equipped to develop a game that suits the demands of their customers, and exploits the intrinsic motivation to participate in social behaviour. Hence, the following main research question is stated:

"What factors contribute to gaming behaviour in Steam's gaming environment?"

To provide a more comprehensive answer to the research question, the main research question is divided into four sub-questions:

Sub-question 1:	What is the main effect of gaming behaviour of an individual?
Sub-question 2:	What characteristics of the relationship between peers, moderate gaming behaviour?
Sub-question 3:	What characteristics of the game moderate gaming behaviour?
Sub-question 4:	What statistical model can achieve the highest performance when predicting peer effects?

Due to the academic properties of this research, a predetermined approach for answering the research question is followed. First of all, the existing literature is consulted and all factors that might be of importance to gaming behaviour are identified in section 2. This will form the basis for the conceptual framework and the proposed hypotheses. Both will be presented in section 2.5. Afterwards, the process of data collection is stated in section 3. In here, an outline of the various sources that are used to consolidate one data set is given and the descriptive statistics are presented. This process is followed by a thoroughly described methodology. In section 4, three different statistical models are described, as well as the methods that assist in achieving the highest performance of these models. The results of these models will be presented in section 5. This section will conclude with a preliminary assessment of the hypotheses and an update of the conceptual framework. Based on these assessments, section 6 will discuss the accepted and rejected hypotheses.

Lastly, in section 7, a comprehensive answer to the research question is given, followed by the limitations of this research and the directions for further research.

1.6 Research structure

Structure				*Contents*
Part 1	Section 1	Theoretical	Introduction	Significance Research Questions
	Section 2	Theoretical	Conceptual framework	Existing literature Conceptual framework
	Section 3	Empirical	Data	Data collection Descriptive statistics
Part 2	Section 4	Empirical	Methodology	Research directions
	Section 5	Empirical	Results	Analysis of findings
Part 3	Section 6	Synthesis	Discussion	Interpretation of findings
	Section 7	Synthesis	Conclusion	Answer to RQ Limitations Directions for further research

Table 1 - Research structure

2. Conceptual framework

The behaviour of individuals can be affected by many factors which can also be observed in a gaming environment. In such an environment, a behavioural change is observed as an increase in in-game time. It is assumed that this change in gaming behaviour is a direct effect of the gaming behaviour of an individuals' peers. This effect is referred to as the main effect (i.e., behavioural change of peers) and is moderated by several variables that differ in their source. Namely, moderators that are based on the relationship characteristics between peers and moderators that are based on the characteristics of a game. The specifics of these factors will be discussed in the following section. First, the main effect is identified, followed by eight moderators: two moderators are based on the relationship characteristics and six moderators are based on the characteristics of a game. Supported by this, nine hypotheses are formulated. After that, the last hypothesis about model performance is stated. For clarity, all hypotheses, as well as a visual representation of the conceptual model is presented in the last part of this section.

2.1 Main effect: behaviour of peers

According to Ryan (2017), "Peer effects refer to externalities in which the actions or characteristics of a reference group affect an individual's behaviour or outcomes". More commonly, peer effects are called peer pressure or peer influence. Where peer pressure is referred to as behaviour

that may provoke or mislead people into doing unnecessary or dangerous things, peer influence is the behaviour of one's peer group that can act as a guideline for people (Panigrahi, 2020). To be more precise, peer pressure usually refers to negative externalities, while peer effects are often positive externalities.

Similar behaviour can also be observed for gamers. Gamers tend to invest additional hours into a game when their peers increase the in-game time of that particular game. The study of Amialchuk and Kotalik (2016) uses grade-level peers (i.e., peers from the same grade in high school) instead of nominated peers (i.e., peers appointed by the individual itself) to assess peer groups. Nominated peer groups lead to groups of close friends that are likely to discuss gaming topics and, therefore, also participate in gaming together. This method creates groups with similar interests and characteristics, whereas a grade-based solution, results in groups where the correlated effects can be separated from the endogenous peer effect. Despite this, nominated peer selection still is an appropriate method for assessing relevant peer groups. In this study, the peer groups are based on the friend list of each player. So, it is assumed that when a player sends a friend request to another player, they have a pre-existing connection. This method can be compared to the nominated peer selection, which is described above. For this research, it is assumed that an individual's gaming behaviour is mainly affected by the gaming behaviour of its peers. Hence, the main effect of gaming behaviour is the gaming behaviour of one's peers. These effects that peers have on one another are called peer effects.

2.2 Moderator effect: relationship characteristics

In this research, the main effect is moderated. This implies that other variables affect the strength of this main effect. These variables can intensify or deteriorate the relationship between two peers (i.e., main effect). This research assumes relationship and game characteristics to be moderators of the relationship between two peers. In the following section, the relationship characteristics are discussed.

2.2.1 Number of peers

Within almost every social structure, there exists an individual or small group of people who are likely to influence other persons in their immediate environment (Katz & Lazarsfeld, 1955). These people are commonly referred to as opinion leaders of a social structure. It is, however, important to note that these opinion leaders are not leaders of the social structure. Opinion leaders are not the head of an organization or whose opinion is exerted through media or authority

structures (Watts & Dodds, 2007). Most literature suggests that opinions leaders, such as influencers, are defined by a high number of social connections, among others (De Veirman, Hudders, & Nelson, 2009). However, the research of Zsolt, Zubcsek, & Sarvary (2011) suggests otherwise. Namely, the average influential power of an opinion leader decreases as their number of friends increases. In this sense, a gamer with many friends has more competing influences than a gamer with few friends. Hence, the number of peers (i.e., the number of friends of a gamer) is identified as a moderator variable.

According to Harrigan, Achananuparp, & Lim (2012), popular individuals - or gamers with many friends – "act as 'inefficient hubs' for social contagion: they have limited attention, are overloaded with inputs, and therefore display limited responsiveness to viral messages". In other words, popular individuals have low influential power on their peers because their opinions are more non-linear and widespread across their network. Thus, it can be stated that the number of peers of an individual, moderate gaming behaviour, such that the effect is stronger when a gamer has fewer peers.

2.2.2 Similar preferences

As described in the research of Wu and Huberman (2007), communities are hard to influence. This is because the individuals inside the community are unresponsive to external opinions. Moreover, the strength of a relationship between two individuals depends on the overlap of one's friendship network (Granovetter, 1973). Complementary, this feature also supports the power of a community. Namely, the shared values of communities have a positive effect on trust in, and the relationship with the community (Wu, Chen, & Chung, 2010). This suggests that individuals within the community are easily influenced by other individuals inside the community because of their similar interests.

According to Cha, Haddadi, Benevenuto, & Gummadi (2010), the most influential individuals have the power to influence other individuals on a variety of topics. However, it is more common that influential individuals are limited to one topic. This supports the findings of Harrigan, Achananuparp, & Lim (2012) as they state that community structures have an increased effect on social contagion. In this manner, the authors contradict the theory that there exists less internal contagion because of "inherent redundancy and lack of novelty within a community". As mentioned before, communities have a higher contagion level because of similar interests and characteristics. With regards to gaming, this might implicate that the relationship between two peers is amplified by the similarities between these gamers. In here, the similarities between gamers are characteristics

such as time invested in a game. Hence, it is stated that similarity between peers moderate gaming behaviour, such that a stronger effect exists when two players are more similar.

2.3 Moderator effect: game characteristics

As mentioned in section 2.2, the behaviour of individuals in a gaming environment can be affected by their peers. In turn, this effect is moderated by the characteristics of the relationship and the game. In the previous section, two moderators of the characteristics of the relationship are discussed and in the following section, six moderators of the characteristics of the game will be discussed. These moderators are characteristics of a game that affect the relationship between two peers for a given game.

2.3.1 Price

Within the gaming industry, building a loyal customer base is highly important. The study of Gummerus, Liljander, Pura, & Van Riel (2004) argues that content-based service providers benefit from this when attracting advertisers and sponsors. Additionally, they pinpoint that "lack of trust is one of the most important reasons for consumers not adopting online services that involve financial exchanges". This loyalty can be exhibited through increased in-game time and increased purchases, among others. One of the main drivers that affect loyalty and consumer purchases, is the price of a game. (Xia, Monroe, & Cox, 2004) (Valvi & West, 2013). The abovementioned concepts are captured in the research of Liao, Tseng, Cheng, & Teng (2020). The writers of this research hypothesize that perceived price fairness is positively related to gamer loyalty. As such, the effect of this moderator is stronger when the price is higher.

2.3.2 Downloadable content (DLC)

Another aspect of price fairness of a game is the amount of content that is included. Back in the days, a consumer bought a game and, when all content is exhausted, the consumer buys a new game. Today, buying that new game has become redundant. If gamers demand new content, more often than not, they can pay to add content to the initial game. This content is new material that surpasses the initial content delivered by the base game. Adding content to a game has become prevalent in the gaming industry (Lee, Jett, & Perti, 2015). Its most common shape is Downloadable Content (DLC) and includes in-game items such as new maps, levels or characters.

Another shape of additional content is microtransactions. These transactions are small financial exchanges to change in-game appearances or to speed up the progress in so-called 'freemium' games (e.g., games that seem free-to-play at first but require small transactions to play the game properly) (TechTerms, n.d.a). Very often, these add-ons can be bought as loot boxes. These random-luck boxes contain a variety of in-game items which can be used to speed up the progress or change in-game appearances. This method for making progress in a game has received a lot of criticism (McCaffrey, 2019). Yet another shape of additional content is the 'Season-pass'. This pass was originally invented to counter the revenue loss of the gaming industry due to the second-hand market. The idea was to supply every new copy of a game with a single-use code to access the online capabilities of a game. When a game is redistributed on the second-hand market, the new owner has to purchase a new code to regain access to the online capabilities of the game (Williams, 2017). This 'Season-pass' received a lot of criticism and was quickly changed into a different pass. This pass either included the future add-ons of a game for a reduced price, or access to small add-ons – such as appearance changes – during the season of a game (Williams, 2017).

Both the price and downloadable content of a game can play a part in increasing in-game time. Price fairness leads to higher customer loyalty which, in turn, leads to more in-game time. Besides that, downloadable content and its addictive properties also influence in-game time positively. Taking this into account, a gamer could either adapt to a game or invest more time in a game. With regards to this research, the amount of downloadable content and price (fairness) might moderate gaming behaviour, such that the effect is stronger when a game has more DLC or a higher price.

2.3.3 Recommendations

In the current digital era, most companies are visible online. Even though a company is not actively building its online image, consumers are creating an online image of a company. Combined with the rapidly increasing speed at which information travels, consumers are highly susceptible to the image of a brand. This image is largely built on the opinions and recommendations of other consumers. According to Senecal and Nantel (2004), less than one fourth (i.e., 22.5%) of consumers buy a product without consulting the recommendations of a product or brand. This suggests that a product with more or higher rated recommendations is bought more often than a product with less or low rated recommendations. Thereby taking into account that the source of the recommendation (i.e., retailer, dependent third-party or independent third-party) and the manipulator of the website (i.e., other consumers, human experts or recommender system) play an

important part. Senecal and Nantel (2004) stipulate that, on average, 30 per cent of the consumers use recommendations written by other consumers on the website of the retailer.

Furthermore, 82 per cent of consumers choose quantity over quality (Senecal & Nantel, 2004). This implicates that a consumer is more likely to choose a product with a higher number of recommendations and a mediocre rating, over a product that has a smaller number of recommendations and a high rating (Storm van Leeuwen, 2019). It is common for gaming environments to have a recommendation system, similar to the ones that retailers use. This system allows gamers (i.e., consumers of the gaming industry) to recommend games to other gamers through written text and a rating system (e.g., thumbs up or thumbs down) (Steam, n.d.a). In that sense, a game with more or better recommendations might be bought more often and is a measure of popularity. In that sense, a gamer is more inclined to buy a game which, in turn, increases the overall in-game time of a gamer for a given game. Thus, it can be stated that recommendations moderate gaming behaviour, such that its effect is stronger when a game has more recommendations.

2.3.4 Achievements

According to Huotari and Hamari (2011), the use of game mechanics has increased rapidly due to evolutions in the game industry. These game mechanics are referred to as "gamification" and are a new set of marketing methods to increase "customer retention and engagement". A more common used gamification is an achievement system (Hamari & Eranti, Framework for Designing and Evaluating Game Achievements, 2011). These systems are similar to customer loyalty cards that are often used in marketing programs. The benefit of these loyalty programs is its focus on retaining customers, which is preferred over gaining new customers (Nunes & Dréze, 2006). The results of Hamari (2014) show that gamification has a positive effect on engagement but are dependent on the context in which it is used. For game environments, gamification mostly has a positive effect on engagement. This is also exemplified by Hamari (2017) which states that gamification increases user engagement through a badge system. This system is part of an achievement system and includes "optional rewards and goals" beyond the main goal of the game. Besides that, a badge consists of a visual and textual element that explain the requirements to obtain the badge. Also, this system is used in most games and proven that it increases intrinsic commitment to a game through challenges (Malone, 1981). Taking the previous into account, this implicates that the in-game time of a gamer might increase if a game includes an achievement

system. Moreover, the number of achievements in the game could moderate gaming behaviour, such that more achievements in a game results in a stronger effect of the moderator.

2.3.5 Genres and online capabilities

As mentioned earlier, the gaming industry has been expanding steadily over the past decade. This expansion has led to the introduction of countless new games, which produced a myriad of genres in the gaming industry. Statista (n.d.c) states that 'action' is the most popular genre, followed by 'shooters', 'role-playing games' (RPGs), 'sports' and 'adventure'. In more recent years, video games with online capabilities have increased in popularity to the extent that 56 per cent of the most frequent gamers play multiplayer games. Moreover, online capabilities influence the decision to purchase a video game in 50 per cent of the cases (The Entertainment Software Association (ESA), 2018).

The research of Lemmens and Hendriks (2016) examines whether Internet Gaming Disorder (IGD) has involvement with the online capabilities and genre of a game. Additionally, they research the addictiveness of nine different genres of games by examining the relationship between IGD and nearly 3,000 games in a sample of 13 to 40-year-olds. Their data shows that the time spent playing online games did not differ significantly from the mean of offline games. However, the results of online and offline capabilities within the game genres in relation to IGD shows different results. Namely, disordered gamers spent more time playing online 'shooters' and online 'RPGs' than non-disordered gamers. At the same time, time spent on offline 'shooters' and offline 'RPGs', did not differ significantly between disordered and non-disordered gamers. This implicates that video games with online capabilities – and within one genre - have more addictive properties than games that do not have online capabilities. Lemmens and Hendriks (2016) state that this is due to the social elements of online games. The urge to interact with other gamers has an impact on the addictive properties, and hence the in-game time of a gamer. This is, however, not a reason to suggest that games without online capabilities are not addictive.

Lemmens and Hendriks (2016), use genres and online capabilities to predict IGD. According to the World Health Organization (2018), IGD includes a pattern where a person gives "increased priority to gaming over other activities to the extent that gaming takes precedence over other interests and daily activities". Implicitly, it can be concluded that IGD and hours spent gaming have a positive relation. This might explain why some games, that vary in genre and online capabilities, lead to more time spent on that particular game than others. In sum, the literature

suggests that the moderator effect of online capabilities is stronger when a game has one of those capabilities. Meanwhile, the moderator effect of the genre of a game is more ambiguous.

2.4 Model performance

Throughout this research, multiple statistical methods will be used. These methods are used for one of two reasons. Namely, for assessing the effect of an independent variable (IV) on a dependent variable (DV) or to predict in what circumstances peer effects occur. In the latter case, a binomial regression model and Random Forest (RF) is used. These statistical methods will be discussed in depth in section 4 and differ in their predicting performance, thereby taking into account that they have no explanatory value for the relationship between a DV and IV. The performance of such models has been the scope of multiple studies. Throughout literature, RFs are largely favoured over regression models in predicting experiments. The research of Couronné, Probst and Boulesteix (2018), who performed a benchmark experiment with 243 real data sets, shows that RF performs better in 69 per cent of the times, compared to a regression model. Muchlinski, Siroky, He and Kocher (2016), as well as Kirasich, Smith and Sadler (2018) suggest that RFs have higher predicting accuracy in binary classification problems. Therefore, based on literature, an RF is assumed as a better performing model, compared to the binomial regression model. Hence, it is hypothesised when predicting peer effects, a random forest can achieve the highest performance.

2.5 Visual representation

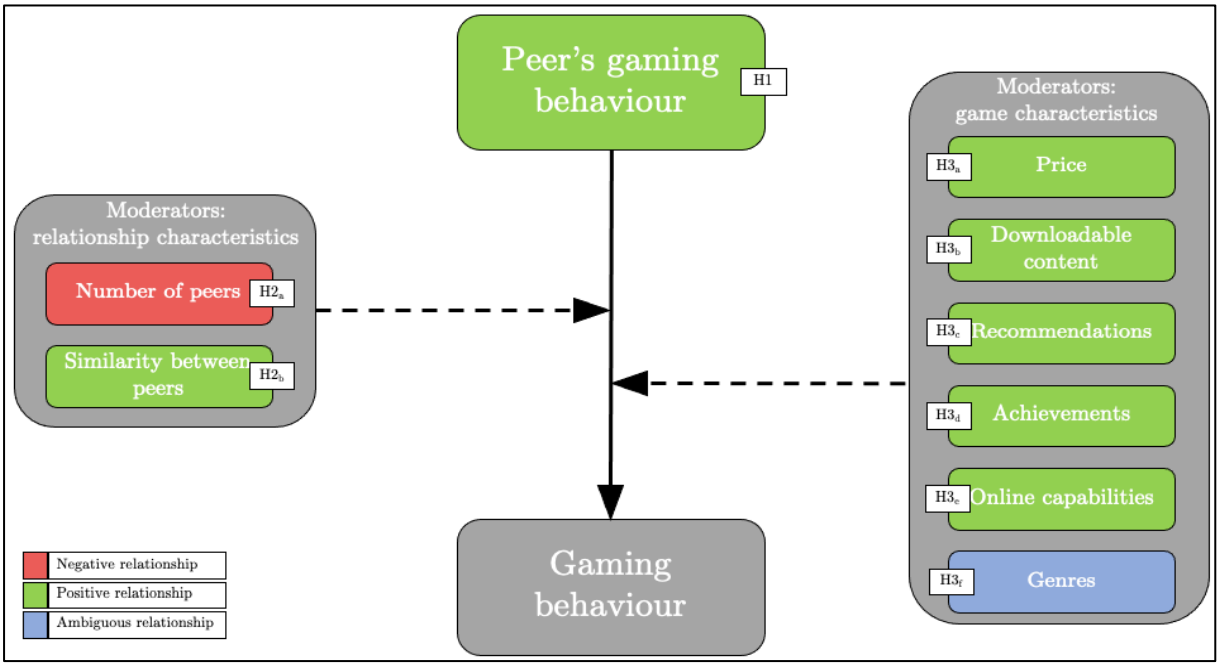

Figure 1 - Conceptual framework

In figure 1, a visual representation of the conceptual framework can be seen. In this research, it is assumed that a peer's gaming behaviour is the main effect of the gaming behaviour of an individual (i.e., H1). The main effect, however, is moderated by the characteristics of the game and the characteristics of the relationship. In total there are two moderators of relationship characteristics. Firstly, the moderator effect of the number of peers on gaming behaviour. Literature suggests that they have a negative relationship (i.e., $H2_a$). This implicates that peer effects decrease if a peer has fewer peers. In turn, this might reduce the in-game time of a gamer for a given game. Secondly, the similarity between peers might have a positive effect on peer effects, indicating that in-game time for a given game might increase if two peers are more similar (i.e., $H2_b$).

Furthermore, the game characteristics consist of six moderators. Firstly, literature suggests that price and downloadable content have a positive effect on in-game time (i.e., $H3_a$ and $H3_b$). These characteristics increase customer loyalty, which increases the likeliness that a gamer is willing to invest more time into a game. Secondly, recommendations also have a positive effect on in-game time (i.e., $H3_c$). Various studies suggest that consumers base their opinion of a product on the opinion of other consumers. This phenomenon also occurs with games and might result in a change in gaming behaviour. In addition to that, an increase in available achievements in a game could result in a rise in the in-game time of a gamer (i.e., $H3_d$). It is suggested in literature that achievements enhance intrinsic motivation, and hence gaming behaviour. Next, online capabilities also increase in-game time due to its social elements. Research indicates that this is related to the urge to interact with other individuals. Lastly, literature cannot pinpoint the exact relation between genres and in-game time. Therefore, genres are assumed to have an ambiguous effect on gaming behaviour. For clarity, all hypotheses are summed up in table 2.

	Hypotheses
H1	"The main effect of gaming behaviour is the gaming behaviour of one's peers (i.e., peer effects)
$H2_a$	"The number of peers moderate gaming behaviour, such that peer influence is stronger if a gamer has fewer friends"
$H2_b$	"The similarity between peers moderates gaming behaviour, such that peer influence is stronger when two players are more similar"
$H3_a$	"The price of a game moderates gaming behaviour, such that peer influence is stronger if a game has a higher price"
$H3_b$	"The amount of available DLC of a game moderates gaming behaviour, such that peer influence is stronger if a game has more available DLC"
$H3_c$	"The number of achievements of a game moderates gaming behaviour, such that peer influence is stronger if a game has more achievements"
$H3_d$	"The number of recommendations of a game moderates gaming behaviour, such that peer influence is stronger if a game has more recommendations"

$H3_e$	"The online capabilities of a game moderates gaming behaviour, such that peer influence is stronger if a game has specific online capabilities"
$H3_f$	"The genre of a game moderates gaming behaviour, such that peer influence is stronger for specific genres"
H4	"When predicting peer effects, a random forest can achieve the highest performance"

Table 2 - Overview of hypotheses

3. Data

In the following section, the data will be thoroughly described. Firstly, the data source is explained, followed by the process of data collection. This process is separated into a description of the raw data and a transformation of the useful data. Afterwards, it is demonstrated how the data is combined with data from a different source and how several variables are created. This results in a data set with many different variables and, for clarity, all variables are described in detail in section 3.4. The section concludes with the descriptive statistics and distributions of the variables.

3.1 Data collection

As mentioned earlier, all data in this study is gathered from the Steam databases. However, all crucial player data is collected from a Steam database that is previously used in the study of O'Neill, Vaziripour, Wu, & Zappala (2016). Furthermore, the data that contains all friendships, is collected using the Application Programming Interfaces (APIs) of the Steam database (i.e., Steam Web API).

3.1.1 Raw data

The Steam database of O'Neill et al. (2016) contains valuable data about all players that have ever registered to Steam. This is raw data such as *Steam ID, App ID* and *playtime forever*. A Steam ID is a unique player ID assigned to a player after registering to the Steam network. This ID is represented in multiple different ways, however, only the 64-bit representation is used for this study. An example of a Steam ID is "76561198019607437". A Steam ID always begins with the same seven digits (i.e., "7656119") and is, from there on, constructed sequentially (O'Neill, Vaziripour, Wu, & Zappala, 2016). Further, the *App ID* is a unique ID given to one game only. According to Steam (n.d.b), "a single product ID will not span multiple applications", implying that every variation of a game has another unique ID. With regard to the in-game time of gamers, the database contains the variable *playtime forever*. This is the total in-game time of a gamer for a given game,

up to the data collection, measured in minutes. Besides that, the database contains two different data collections: The first data collection is from June 11th, 2013 through June 25th, 2013 and the second data collection is from August 1st, 2014 through August 14th, 2014. Using the difference between the first data collection and the second data collection, it is possible to calculate the increase in in-game time between two data collections (O'Neill, Vaziripour, Wu, & Zappala, 2016). From here on, the first collection will be referred to as $t = 0$ and the second collection as $t = 1$.

3.1.2 Players and peers

The database, as constructed by O'Neill, Vaziripour, Wu, & Zappala (2016), contains data of all 108.7 million players that were registered in June 2013. The sheer size of this database made the research too computationally intensive. Therefore, a random sample of 5,195 unique players and 125,134 observations is used for further examination. This group of players is referred to as the 'players'. In order to collect all peers of the players, the Steam Web APIs are used. APIs are a collection of web addresses that can be used to establish a connection between an application and a partition of another application, such as a database. In this case, the API makes a connection between a local programming environment (i.e., RStudio) and the database of Steam that resides the peer lists of all players. This API collection resulted in 107,908 peers, of which 107,615 are unique *Steam IDs*. This inequality can be explained by the fact that 293 players in this collection are a peer of more than one of the players. The peer list contains all peers of the players, and these individuals will be referred to as the 'peers'. As mentioned before, in total there are 107,615 unique *Steam IDs* in this list. Using the database of O'Neill et al. (2016), the *playtime forever* at $t = 0$ and $t = 1$ of the peers is collected. This resulted in a data set containing 2,558,7655 observations of 107,615 unique *Steam IDs*. In table 3 an overview of the different data sets can be seen.

Data set	*Player data*	*Peer list*	*Peer data*
Description	Data of the players.	Specifies friendships between players	Data of all peers
Unique players	5,195	107,615	107,614
Observations	125,134	107,908	2,558,765
Source	(O'Neill, Vaziripour, Wu, & Zappala, 2016)	(Steam, n.d.d)	(O'Neill, Vaziripour, Wu, & Zappala, 2016)

Table 3 - Overview of data sets

3.2 Combining databases

After completing the last data collection, three different data set are created (i.e., player data, peer list and peer data) (table 3). During the preparation of the data for the analysis, one database is compiled out of the three different data sets that are mentioned in table 3. The newly compiled database is a consolidation of the player data and the peer data through the peer list. Because this research is focussed on peer effects, measured between peers that play the same game, only these observations are required for this analysis. Thus, the new data set only includes observations of players and peers that have a match based on *Steam ID* and *App ID*. This merge between the data sets follows the opposite direction of the data collections. This implies that first, the peers' data set is merged with the peer list based on the *Steam ID*. Then, the data set is merged with the players' data set, based on *Steam ID* and *App ID*. In this manner, only observations where players and peers play games with equal *App IDs*, remain in the data set.

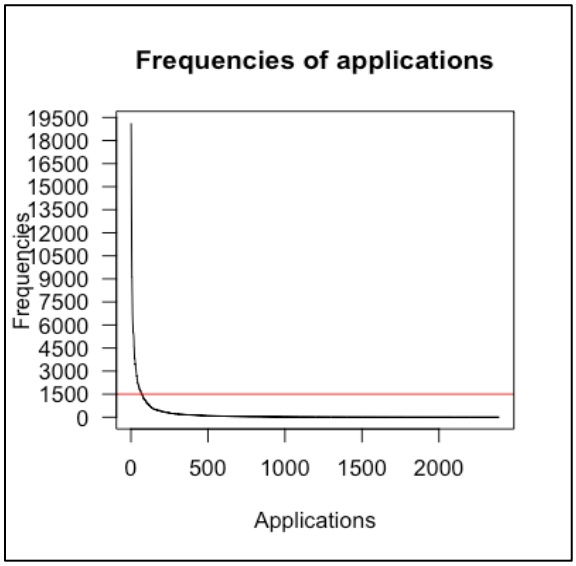

Figure 2 – Initial situation of application frequencies

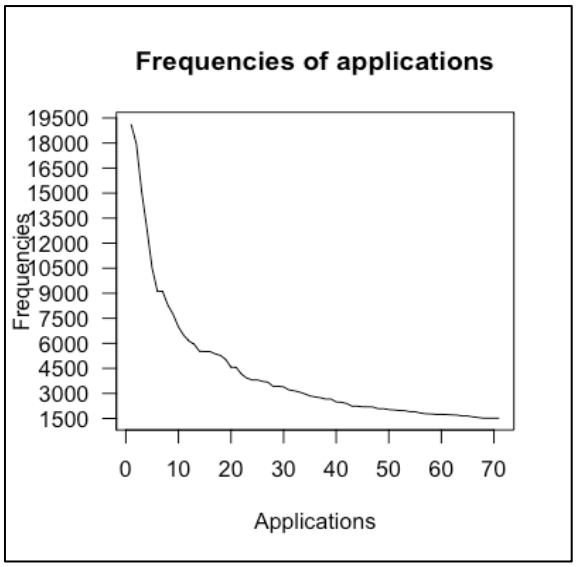

Figure 3 - Ultimate situation of application frequencies

3.3 Transformations

After the data sets are merged, all observations in the new data set are a friendship between two players for a given *App ID*. Additionally, the *playtime forever* of the players and peers of the data collection in $t = 0$ and $t = 1$ are given. In the process of cleaning the data, two transformations are performed. Firstly, observations where the *playtime forever* of the players in $t = 1$ is Not Available (NA), are deleted from the data. In doing so, only observations where the player and peer owned the game at $t = 0$ and $t = 1$, or owned it only at $t = 1$, remained in the data set. The latter case occurs when, for instance, a game is released after the collection in $t = 0$. In this case, the

playtime forever in $t = 0$ is NA as well. These NAs are transformed into zero, indicating that a player has not played the game yet. Afterwards, the data set contained 480,611 observations. However, some of the *App IDs* in these observations occur only a few times which might yield less robust results. Therefore, only observations of *App IDs* that occur more than 1500 times are included in the data set. The threshold of 1500 is chosen arbitrary and decreased the number of observations to 260,389. In figure 2 and 3, the initial data set, and the subset of the data set can be seen respectively.

3.4 Variables

3.4.1 Dependent variables

The first variable is created by merging the variables *playtime forever players* at $t = 0$ and *playtime forever players* at $t = 1$. This new variable indicates the increase in in-game time of players between the data collection at $t = 0$ and $t = 1$, measured in minutes. It is calculated by subtracting *playtime forever* at $t = 0$ from *playtime forever* at $t = 1$. The variable is called *playtime increase player* and will be used as continuous dependent variable (DV) of the linear regression model. Besides that, the binomial regression and Random Forest model in this research require a discrete DV. For those models, a binary counterpart of *playtime increase player* is created. This variable is called *playtime binary* and assumes the value '1' if peer effects are present and assumes the value '0' if this is not the case. In this research, *playtime binary* assumes the value '1' if *playtime increase player* is higher than zero. If it is equal to zero, then *playtime binary* assumes a value of '0'. This results in 144,235 zeros and 116,154 ones.

3.4.2 Relationship variables

Furthermore, three variables, related to the characteristics of the relationship between two peers, are created. Firstly, a variable that determines the similarity between players is created. In this case, the similarity is calculated using the Euclidean distance metric. This metric can be used to calculate the similarity (i.e., Euclidean distance) between the vectors q_i and p_i in a $p * n$ matrix $\boldsymbol{Z}$ and can be seen in formula (1).

$$d_{pq} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \qquad (1)$$

In here, d_{pq} is the Euclidean distance between the vectors q_i and p_i in matrix $\mathbf{Z}$. This can be computed for all i elements and results in a dissimilarity matrix of Euclidean distances $\mathbf{D}$ (Elmore & Richman, 2001). With regard to the data, q_i and p_i denote the *playtime forever* of *App ID* i in $\mathsf{t} = 0$ that player q and peer p have in common. This implicates that one similarity score is calculated for a player and peer, based on their respective *playtime forever* of all *App IDs* they have in common in $\mathsf{t} = 0$. For this computation, the *playtime forever* in $\mathsf{t} = 0$ is scaled in such a manner that it ranges from 0 to 1. In doing so, the scale of the Euclidean distance ranges from 0 to 27 where a value close to zero indicates highly similar players and close to 27 indicates highly dissimilar players. Because scaling cases - where the *playtime forever* for all *App IDs* is zero - did not yield numeric results, observations, where the Euclidean distance resulted in NA, are transformed to the mean of all Euclidean distances.

Secondly, a variable is constructed that indicates the size of a player's social network size. This variable is called *number of peers* and denotes the number of peers a player has. Lastly, the variable *playtime increase peer* is constructed. This variable is created in the same manner as *playtime increase player* and indicates the increase in in-game time of a peer for a specific game, measured in minutes.

3.4.3 Game variables

In addition to the variables that illustrate the characteristics of the relationship between two players, a set of variables is collected that contains statistics about all *App IDs.* These variables are collected through an API that establishes a connection with the Steam Storefront instead of the Steam Web APIs. This API collects data from the Steam Store and includes the following variables: *Name*, *Price*, (number of) *recommendations*, (number of) *achievements*, (number of) *DLCs*, binary variables for 39 different *categories* and binary variables for 24 different *genres*. The API is limited in such a manner that it only collects data from applications that are played by the peers in $t = 1$. This resulted in the collection of 68 different variables for 4,179 different games.

The number of different variables did, however, decrease after some of the *categories* and *genres* are deleted. Firstly, all *categories* and *genres* that do not have a single positive observation (i.e., the value 'one'; indicating that the *App ID* manifests the specific *category* or *genre*), are deleted from the data. Secondly, all remaining *categories* and *genres* that do not contain any valuable information for this research are also deleted from the data. The *categories* that remain in the data set are all related to the online capabilities of a game. Therefore, from now on, *categories*

will be referred to *online capabilities*. In table 12 and 13 in appendix A, an overview of the variables can be seen.

3.5 Descriptive statistics – Continuous variables

To derive meaningful insights from the data, 22 variables are included in the statistic models. Of them, eight are continuous variables. In table 4, the descriptive statistics of these variables can be seen. Firstly, *playtime increase player* ranges from 0 to 221,263 indicating that at least one player in this population did not increase its playtime between $t = 0$ and $t = 1$, and at least one player increased its playtime by 221,263 between $t = 0$ and $t = 1$. On average, players increase their playtime by 2,473.62, while the median is 0.00. This implicates that most observations assume a value of 0.00 and the distribution highly skewed to the right. The standard deviation of *playtime increase player* equals 11,821.73. Using the coefficient of variation (CV) (i.e., standard deviation divided by the mean), this indicates that the values of this variable are relatively dispersed across its range.

Secondly, *playtime increase peer* has, compared to *playtime increase player*, a greater range: the lowest value is 0 and the highest is 610,616. Equal to *playtime increase player*, the median of *playtime increase peer* is 0.00, which is lower than the mean. This implies that most observations assume a value of 0.00 and that the distribution is right-skewed. The standard deviation of *playtime increase peer* is slightly higher than the playtime of players, indicating that the values are more dispersed from the mean. Thirdly, the lowest *price* of games is 0.00 (i.e., a game that is free-to-play) and the highest *price* is 59.99. On average the *price* of a game is 10.04 and the median is 8.19, suggesting a distribution that is slightly skewed to the left. The standard deviation of *price* is 12.29, which is lower compared to the previously discussed variables and indicates that the values are more clustered around the mean.

Further, *number of peers* range from 1 to 1,045 and on average a player has 105.30 peers. The median of this variable is relatively close to the mean, which implicates that it has a more symmetrical distribution. Besides that, the standard deviation of this variable is 145.51, which is evidence of a distribution where the values are more clustered around the mean. In addition to that, *Euclidean distance* ranges from 0.00 to 27.35, which is a relatively large range. Also, the mean is equal to the median, indicating that the distribution is perfectly symmetrical. The standard deviation of this variable is 3.08, which implies that the values are less clustered around the mean and more spread out across the distribution.

Moreover, the number of *recommendations* range from 332 to 2,817,294 and the mean is 163,032.07. Because the median is severely lower than the mean, this variable has a non-symmetrical distribution. *Recommendations* has a standard deviation of 517,169.24, which suggests that the values are more dispersed across the distribution. The *achievements* of a game range from 0 to 1,207 and a game has 99.45 *recommendations* on average. The median of this variable is relatively close to the mean, which suggests that the distribution is more symmetrical. Additionally, *achievements* has a standard deviation of 166.71, implying that the values are more clustered around the mean. Lastly, the number of *DLCs* of a game range from 0 to 60 and a game has 4.56 DLCs on average. This variable is slightly skewed to the right and the values have a high level of dispersion across the distribution.

N = 260,389	*Mean*	*St. dev.*	*Median*	*Min*	*Max*
Playtime increase player	2,473.62	11,821.73	0.00	0.00	221,263.00
Playtime increase peer	3,262.38	14,953.96	0.00	0.00	610,616.00
Price	10.04	12.29	8.19	0.00	59.99
Number of peers	105.30	145.51	63.00	1.00	1,045.00
Euclidean distance	3.87	3.08	3.87	0.00	27.35
Recommendations	163,032.07	517,169.24	15,326.00	331	2,817,294.00
Achievements	99.45	99.45	50.00	0.00	1,207.00
DLC	4.56	4.56	1.00	0.00	60.00

Table 4 - Descriptive statistics of continuous variables

3.5.1 Distributions

In figure 4, the distributions of continuous variables are shown. This visual representation confirms several suggestions, made in the previous section, about the shape of the distributions. The x-axis denotes the value of the respective variable and the y-axis denotes the concentration of observations. Essentially, this is a smoothened version of a histogram plot. It shows that p*laytime increase player* and *playtime increase peer* both have severely right-skewed distributions, and most observations assume a value of 0.00. Similarly, many observations of *price* assume a value of 0.00 (i.e., *free-to-play* games). However, this variable also has observations that cluster more around the mean. This results in a distribution that is slightly skewed to the right. In addition to that, *number of peers* is more symmetrical which results in a normally distributed variable. According to the previous section, *Euclidean distance* should also follow a normal distribution. However, due to outliers on the right side of the distribution, the variable is highly skewed to the right. Furthermore, the distributions of *recommendations* and *achievements* are also right-skewed and have many observations that assume a value of 0.00. Similarly, also has a distribution that is skewed to the right.

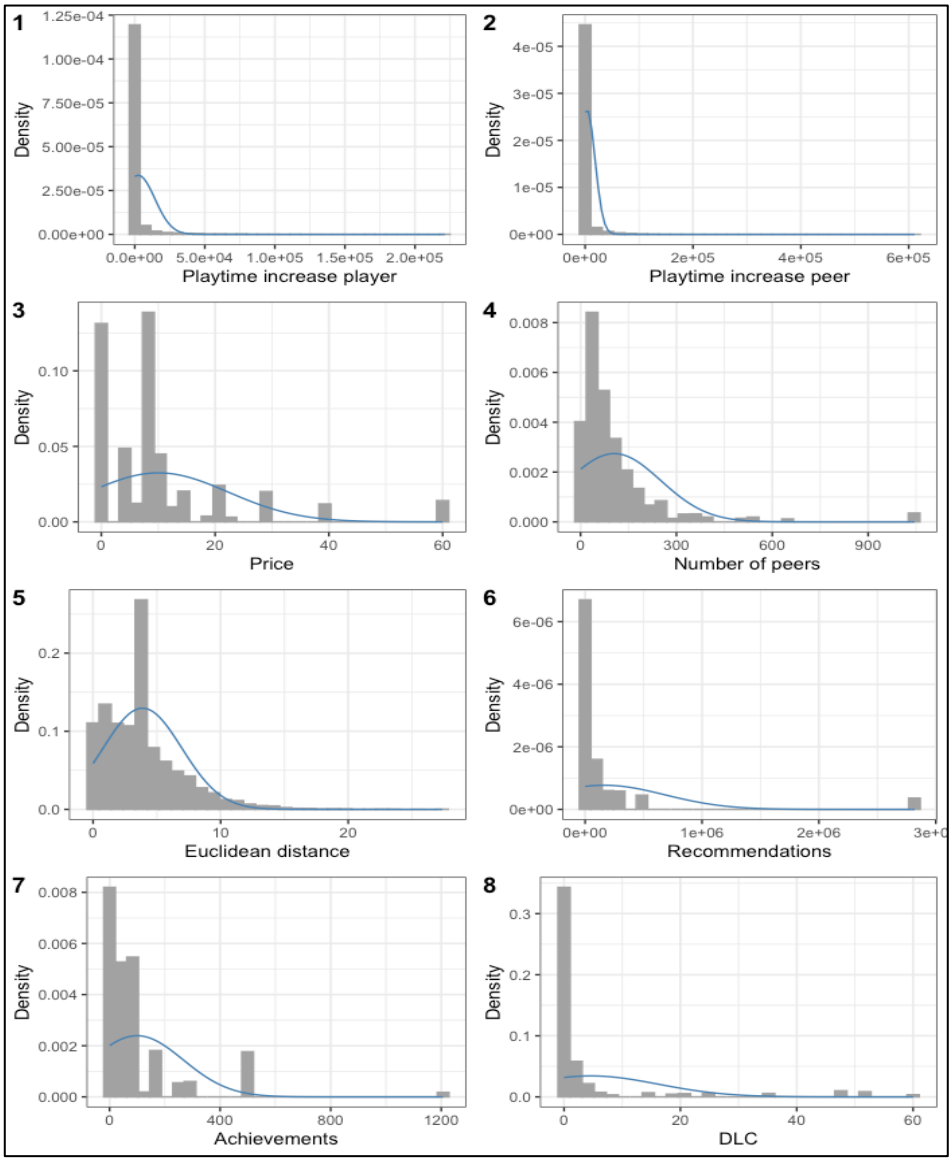

Figure 4 - Distribution plots of continuous variables

3.5.2 Correlations

In figure 5, the correlation between all pairs of continuous variables can be seen. The correlation metric ranges from -1 to 1, where a value of -1 indicates a perfectly negative relation and a value of 1 indicates a perfectly positive correlation All relatively strong correlations between variables are positive correlations. This implies that these pairs of variables are positively related and move in the same direction. Besides that, the strongest correlation is between *price* and *DLCs*. This correlation is estimated at 0.65, which indicates that *price* and *DLC* have a strong relationship that moves in the same direction. At 0.37, the correlation between *playtime increase player* and *playtime increase pear* is the second largest. Consequently, the correlation between both variables and the remaining variables is alike. This suggests that players and peers react similarly to changes

in the characteristics of the relationship and the game. In addition to that, the correlation between *achievements* and *DLC* is 0.23. The remaining correlations are below an absolute value of 0.15.

To ensure that no explanatory variable can be explained by another explanatory variable, it is essential to test for multicollinearity. A high correlation between two explanatory variables is a strong sign of multicollinearity. In literature, a rule-of-thumb for measuring multicollinearity is a correlation that exceeds an absolute value of 0.7. However, no correlation between variables exceeds 0.7. Therefore, based on correlation, no variables are removed from the data due to multicollinearity.

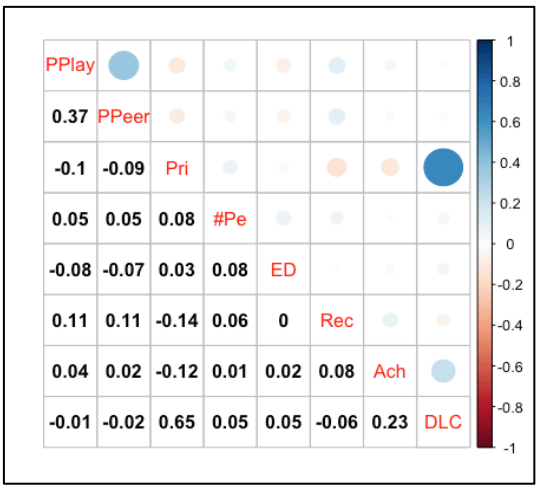

Figure 5 - Correlations of continuous variables

3.6 Descriptive statistics – Discrete variables

As mentioned, this research contains 22 variables. Eight of them are continuous and the remaining 14 variables are discrete. The discrete variables are divided into two categories. First of all, the variables that categorize a game based on its genre. Table 5 shows that games in this research are predominantly labelled as an *action* game, whereas the genres *adventure, indie, RPG, free-to-play, strategy and simulation* occur less frequently. Thereby taking into account that a game belongs to, at least, one genre and it can belong to more than one genre. This resulted in several combinations of genres. The most occurring combinations of genres can be seen in figures 30 in appendix B.

N = 260,389	*Count*	*Percentage*
Action	242,569	93.16
Adventure	33,661	12.93
Indie	35,047	13.46
RPG	35,349	13.58
Free-to-play	45,848	17.61
Strategy	28,729	11.03
Simulation	10,740	4.12

Table 5 – Distributions of genres

Secondly, the variables that are categorised based on their online capabilities. These variables can be seen in table 6. It shows that over 80 per cent of the games are *multiplayer* games and a substantial amount of the games is *co-op*. The remaining online capabilities occur considerably less frequent. Similar to the genre of a game, games can have more than one online capability. This implicates that several combinations of online capabilities exist. The most occurring combinations can be seen in figure 31 in appendix B.

N = 260,389	*Count*	*Percentage*
Multiplayer	209,907	80.61
PvP	51,308	19.70
Co-op	108,563	41.69
Online Coop	37,657	14.46
Cross-Platform Multiplayer	44,165	16.96
LAN PvP	2,469	0.95
Shared/split-screen	26,782	10.29

Table 6 - Distributions of online capabilities

4. Methods

In order to answer the research question, several statistical models are used. The models that are used are discussed in this section. First, a linear regression model is discussed. This model will examine the effect of multiple predictors on a continuous DV. Then, a Generalised Linear Model (GLM) is examined, which will answer a classification problem. This implicates that the DV is transformed from a continuous variable into a binary variable. For this model, the Akaike information criteria (AIC) and Bayesian information criteria (BIC) are reviewed. These criteria can balance the fit of a model with the number of predictors. Next, an RF is considered. This model also answers a classification problem and is used to predict a binary DV. Furthermore, the methods of optimising the hyperparameters of the predictive models are studied according to the Receiver operating characteristics (ROC) curve and Out-of-bag (OOB) error optimization. Lastly, the method of assessing the performance of the models is discussed.

4.1 Linear Regression model

As mentioned before, a linear regression model is used to determine how the IVs affect the DV. Because multiple IVs are used to determine a single continuous variable, the model is called a multiple linear regression. This model uses the least-squared principle to fit a linear function to the data. The linear function of a linear regression with multiple predictors looks as follows:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + \varepsilon_i \quad (2)$$

In formula 2, $\hat{y}_i$ is the predicted response of the i th case, b_0 is the intercept, $b_{1\ldots p}$ are the coefficients of predictors x_i for $i = 1, \ldots, n$, the number of predictors is denoted by p, and ε_i represents the error term.

In this research, the linear regression uses *playtime increase player* as DV and *playtime increase peer* as IV. *Price, number of peers, Euclidean distance*, (number of) *recommendations*, (number of) *achievements* and (number of) *DLCs* are moderators. Therefore, these variables are included as interaction terms of the IV. Besides that, the linear regression includes all available binary variables of *online capabilities* and *genres* as interaction term (table 13; appendix A). This leads to the following linear function (formula 3):

$$\begin{aligned} &Playtime\ increase\ player \\ &\quad = b_0 + playtime\ increase\ peer(b_1 + price * b_2 \\ &\quad + number\ of\ peers * b_3 + Euclidean\ distance * b_4 \\ &\quad + recommendations * b_5 + achievements * b_6 + DLC * b_6 \\ &\quad + online\ capabilities_c * b_c + genres_g * b_g) + \varepsilon \end{aligned} \quad (3)$$

In formula 3, c and g denote the different *online capabilities* and *genres* that are included in the database. The coefficients (b) of the IVs in this linear regression are an estimate of the effect of the IVs on the DV. In turn, these coefficients will help to answer the research question by testing hypotheses one to three.

4.2 Binomial Regression model – classification

If the response variable in a regression model follows a non-linear distribution, a binomial regression is required. This model aims to fit the data to a binomial distribution, which is the number of times an event occurs in n independent Bernoulli trials. In here, p denominates the probability of an event occurring and q is the probability of an event not occurring (formula 4 and 5) (Uspensky, 1937).

$$0 \leq p \leq 1 \quad (4)$$

$$q = 1 - p \quad (5)$$

A binomial regression with a binary response variable is essentially a binary regression with $n = 1$ Bernoulli trials. This type of regression is also known as the logit model (i.e., logistic regression) and requires a discrete DV. This implies that the response variable can only assume a limited number of outcomes where all outcomes in between those outcomes have no meaning. For a binary DV, the outcomes are limited to 0 or 1 (i.e., failure and success). In that case, the mean of the response is the probability p that an event occurs, based on one or more predictors. Otherwise stated, based on several predictors, the logit model predicts the odds that the response variable assumes the value 'one'. This prediction is calculated using 'log-odds', which is the logarithmic outcome of $p/(1-p)$, and models it as a linear combination of all IVs (Moore, McCabe, Alwan, Craig, & Duckworth, 2011). In sum, the logit model looks as follows (formula 6):

$$\log(\frac{p}{1-p}) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + \varepsilon_i \tag{6}$$

The model that will be used in this research for the binary classification is, in terms of IVs, similar to the linear regression that can be seen in formula 7. That being said, the DV is the binary counterpart of *playtime increase player* and is named *playtime binary*. This results in the following formula (formula 7):

$$\begin{aligned} \log\left(\frac{p}{1-p}\right) = b_0 + playtime\ increase\ peer(b_1 + price * b_2 + number\ of\ peers \\ * b_3 + Euclidean\ distance * b_4 + recommendations * b_5 \\ + achievements * b_6\ + DLC * b_6\ + online\ capabilities_c * b_c \\ + genres_g * b_g) + \varepsilon \end{aligned} \tag{7}$$

In formula 7, c and g denote the different *online capabilities* and *genres* that are included in the data. The coefficients (b) of the IVs in this logit regression are associated with the probability p that *playtime binary* assumes a value of '1'. The combination of IVs that results in the best fit will be discussed in section 4.3.

4.3 Predictor selection

Adding all available predictors in the data set to a regression does not necessarily lead to the most accurate model. In some cases, a model risks overfitting when all predictors are added to the linear regression. This predicament occurs in models that are highly dependent on the specific

data that is used to create the model. In most cases, overfitting causes the model to lose accuracy on out of sample data (Webb, 2011). In order to prevent this, a predictor selection method is used. A predictor selection method aims to find a combination of IVs (i.e., predictors) that lead to the highest possible quality of a model, given the available data set. There are different methods for determining the quality of a model. However, when the model includes more than one IV, OOB data is not available and there is a large number of different variations of the model possible, information criteria are good alternatives (Akaike, 1976).

4.3.1 Akaike information criterion

One of the predictor selection methods used in this research is the Akaike information criterion (AIC) (Akaike, 1973). This method can estimate the error of OOB predictions. In estimating the error, AIC can determine the relative quality of a statistical model compared to other models. Thus, AIC is a metric for determining the best set of predictors. AIC is a log-likelihood function with a penalty term and, therefore, belongs to the family of the penalized-likelihood criteria. This family of functions evaluates the fit of the model by adding a penalty term for the complexity of the model. The general function of this family includes a goodness-of-fit term and an overfitting penalty (Formula 8) (Dziak, Coffman, Lanza, & Li, 2012).

$$IC = A_n p - 2\ln(l) \tag{8}$$

In here, A_n is an overfitting penalty for sample size n and p is the number of predictors in the model. Further, l is the log-likelihood (i.e., goodness-of-fit term) which is a metric that determines how likely the observed data is. The model that yields the lowest value for the penalty term minus the log-likelihood has the best relative goodness-of-fit (i.e., a lower AIC yields the highest goodness-of-fit). For AIC, the overfitting penalty is $A_n = 2$, which results in the following formula:

$$AIC = 2p - 2\ln(l) \tag{9}$$

4.3.2 Bayesian information criteria

Even tough AIC is a proper method for determining the goodness-of-fit of a model, predictor selection is not necessarily based on only one method. Kuha (2004) argues that the optimal model can be chosen based on information from more than one criterion. The author suggests a combination of AIC and Bayesian information criteria (BIC) for predictor selection. Just as AIC, BIC belongs to the family of the penalized-likelihood criteria and follows the assumption that a

lower BIC yields the highest goodness-of-fit. This implicates that it is based on the same general function (formula 8), however, with a slight adjustment to the overfitting penalty. Namely, instead of a penalty of $A_n = 2$, the penalty is $A_n = \ln(n)$. This results in the following formula for BIC (Dziak et al., 2012):

$$BIC = \ln(n)p - 2\ln(l) \tag{10}$$

The main difference between the two metrics is what they define as a "good model". On the one hand, AIC is a better metric when the sample size is large, however, tends to favour larger models over smaller models. BIC on the other hand is better in identifying the true model and more consistent for predictor selection because it assigns a higher penalty to more complex models (Kuha, 2004). This implicates that BIC might choose a model that is too small. Therefore, a combination of the two might lead to an even more justified model selection. Thereby taking into account that the chosen model is the model that is highly favoured by both metrics.

4.3.3 Direction of information criteria

Additionally, a backward/forward (i.e., bidirectional) selection method is used to determine the optimal AIC and BIC. This method is a combination of the backward selection method and forward selection method. Backward selection starts with all predictors in the model and sequentially removes the predictors that do not contribute to a better AIC or BIC, whereas forward selection starts with none of the predictors and sequentially adds predictors that contribute most to the AIC or BIC. A combination of the two results in a method that starts with none of the predictors and adds the most contributing predictors. Then, it sequentially removes the variables that do not lead to an improvement of the model (Derksen & Keselman, 1992)

4.4 Random forest model – classification

The earliest mention of the Random Forest model in literature is by Ho (1995). He argued that the decision trees (Breiman, Friedman, Olshen, & Stone, 1984) where RFs are derived from, cannot be grown to full complexity. This is because the level of accuracy decreases significantly when OOB data is used. To prevent this from occurring, Ho (1995) proposed a "stochastic discrimination" method that grows "multiple trees with randomly selected subspaces of the feature space". This implies that for every tree, a random subset of predictors is chosen to grow the tree. In this manner, the RF also performs adequate on OOB data, albeit at the loss of interpretability.

The theory of Ho (1995) is later extended and registered by Breiman (2001) and is based on the bagging principle of Breiman (1996).

4.4.1 Classification and regression trees

As mentioned, RFs are based on decision trees. Breiman et al. (1984) named them Classification and Regression Trees (CART). They are built by randomly selecting IVs and splits until the optimal tree is grown with decision nodes and leaf nodes (figure 32 in appendix C). A decision tree starts with the root node and every node that is followed by another split is called a decision node. If a node is not followed by another split, it is called a leaf node. A CART uses recursive binary splitting to determine the optimal tree. It is a method that evaluates the cost of each split by calculating what split results in the lowest cost and is calculated through a cost-function. There are many examples of cost-functions, however, for a classification problem, the Gini-index is used. This index measures the probability that a randomly chosen variable is not classified accurate. It is also known as an impurity measure because if all objects are randomly distributed between the classes, the Gini-index is 1 (i.e., 100% and impure) which is perfect inequality. On the contrary, the Gini-index is 0 (i.e., 0% and pure) when all objects belong to one class (Gini, 1912) (Ceriani & Verme, 2012). The formula of the Gini-index can be seen in formula 11. In this formula, p_i is the probability that a randomly chosen variable i is classified to a certain class. At each node, the classification decision tree chooses the predictor with the lowest Gini-index.

$$G = \sum_{i=1}^{n} (p_i)^2 \tag{11}$$

4.4.2 Random Forest

An RF is a parallel ensemble method that combines multiple, independently build, decisions trees into one RF (figure 6). In this manner, it can combine many different models with high complexity and low bias into one model with low variance. The advantage of an RF is that the correlation between the models is low. This is because the trees are built on many different data sets. Preferably, an RF is built on B independent training sets (i.e., $\hat{f}_1(x), \hat{f}_2(x), \dots, \hat{f}_B(x)$). This results in the following average prediction function (formula 12):

$$\hat{f}_{avg} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b \tag{12}$$

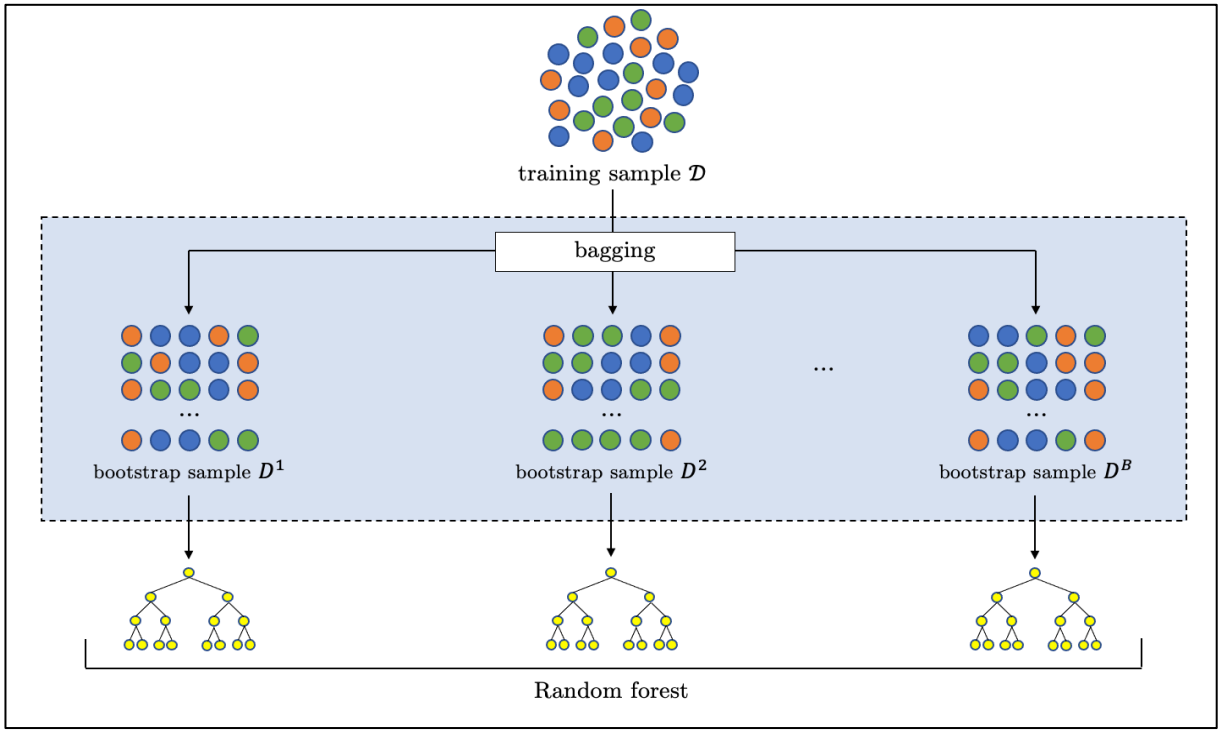

Figure 6 - Overview of Random Forest

4.4.3 Hyperparameters

However, in most cases, it is not possible to use B independent training sets. As an alternative, a bootstrap aggregating method – also known as bagging – can be used. This method, as suggested by Breiman (1996), creates multiple "bootstrap replicates" of the training set and uses these as independent training sets. These replicated data sets are constructed with replacement which implicates that a single observation can occur not at all, once or more than once. Subsequently, when B equally sized bootstrap samples $D^1, \dots, D^B$ are drawn from the original data set $\mathcal{D}$, the bagged decision function can be depicted as (formula 13). According to literature, $\hat{f}_{bag}$ performs similarly as $\hat{f}_{avg}$, albeit with a smaller variance reduction compared to B independent training samples (Rosenberg, 2017). In here, B is equal to the number of iterated decision trees in an RF. Therefore, B is a hyperparameter and will be discussed in section 4.6.2.

$$\hat{f}_{bag} = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b \qquad (13)$$

Besides that, RFs restricts the number of splitting variables that can be chosen randomly, to m features. A rule-of-thumb is to choose $m \sim \sqrt{p}$, where p is the number of features (Rosenberg, 2017). This rule-of-thumb, however, can be optimised to further extend and is also known as a

hyperparameter. In this research, the RF uses the same set of IVs that are used in the linear regression (section 4.1). Considering that it is an ensemble method with m random features, it is not possible to indicate the exact set of IVs that are used in the final RF. Additionally, *playtime binary* is used as DV since this is a classification problem with a binary outcome.

4.5 Transformations

In addition to the transformations in section 3.3, three more alterations will be discussed. Firstly, all continuous variables have been transformed to a logarithmic scale. This transformation ensures that all continuous variables are more conform and are equally valued in the models. Also, logarithmic scales produce a more normally distributed data set as it removes skewness from the variables.

Secondly, the data set is split into a training and test data set. The training set will be used to create the models and the test set contains all out-of-sample observations and is used to assess the performance of the models. Normally, an RF does not require a test set to assess performance, however, for means of comparison, the RF still uses a test set to assess its performance. This will be discussed in depth in paragraph 4.7. During the separation of the data set, 70 per cent of the 260,389 observations is randomly assigned to the training data set and the remaining 30 per cent is assigned to the test data set. In doing so, observations of one particular player may be assigned to both the training and test data set.

Lastly, the binary value *playtime binary* is redistributed because it assumes the value 1 only in a third of the observations which results in an unbalanced data set. This decreases the performance of a model because it will emphasize the majority class and ignores the minority class. It can be corrected through undersampling or oversampling. Undersampling decreases the number of majority class observations and oversampling replicates observations of the minority class to balance the data set. These two techniques, however, raise a new problem; undersampling leads to loss of possible critical information and oversampling adds duplicated data which lead to overfitting. Random OverSampling Examples (ROSE) solves both problems by generating new artificial data from the classes and is based on a "bootstrap form of re-sampling from data" (Menardi & Torelli, 2012). The main goal of ROSE is to combine undersampling and oversampling by generating artificial data that is closely related to the minority class. In this manner, no information is lost, and overfitting is less likely to occur.

4.6 Hyperparameter optimisation

Both models that are used for classification problems (i.e., binomial regression model and random forest) have internal parameters that can be altered. If altering these parameters results in better performing models, it is called hyperparameter optimisation. The hyperparameter that will be optimised in the binomial regression model is the Receiver Operating Characteristics (ROC) curve and is a graph of sensitivity as a function of the inverse specificity (i.e., 1 – specificity). Using this graph, the threshold for positive and negative classifications can be determined. Concerning the random forest, two hyperparameters can be optimised. Namely, the number of iterated decision trees (e.g., B) and the size of the set of features (e.g., p) that can be chosen from at each split. Additionally, it can be argued that prediction selection through AIC and BIC as discussed in section 4.3 is also a hyperparameter.

4.6.1 Receiver operating characteristics curve

In binary classification problems, the outcomes are either positive or negative. This results in four test outcomes (figure 7). Firstly, true positive (TP), also known as a "hit", occurs when the classifier predicts a positive value, and the actual value is also positive. Secondly, true negative (TN) occurs when the classifier predicts a negative value, and the actual value is also negative. This is also known as a "rejection". Furthermore, false positive (FP) occurs if the classifier predicts a positive value while this is, in fact, negative and is also known as "false alarm" (i.e., Type I error). Lastly, a false negative (FN) observation occurs when the classifier predicts a negative value, and the actual value is positive. This is referred to as a "miss" (i.e., Type II error).

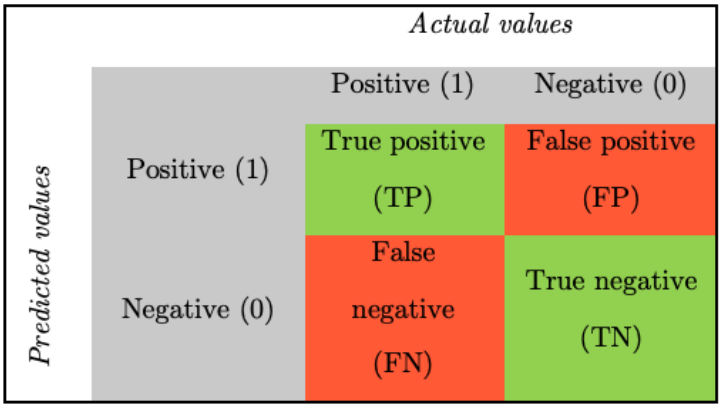

Figure 7 - Classification of outcomes

The ROC curve plots the sensitivity (formula 14) as a function of the inverse specificity (formula 15). Sensitivity and specificity are also referred to as true positive rate (TPR) and true negative rate (TNR) respectively and can both be calculated using TP, TN, FP and FN.

$$Sensitivity = TPR = \frac{TP}{TP + FN} \tag{14}$$

$$Specificity = TNR = \frac{TN}{TN + FP} = 1 - FPR \tag{15}$$

The main purpose of the ROC curve is to determine the threshold of classification that optimises the trade-off between sensitivity and specificity (figure 8). This threshold is the value that has to be exceeded for the classifier the predict a positive outcome. If the predicted value is below the threshold, the classifier predicts a negative outcome (Fawcett, 2006).

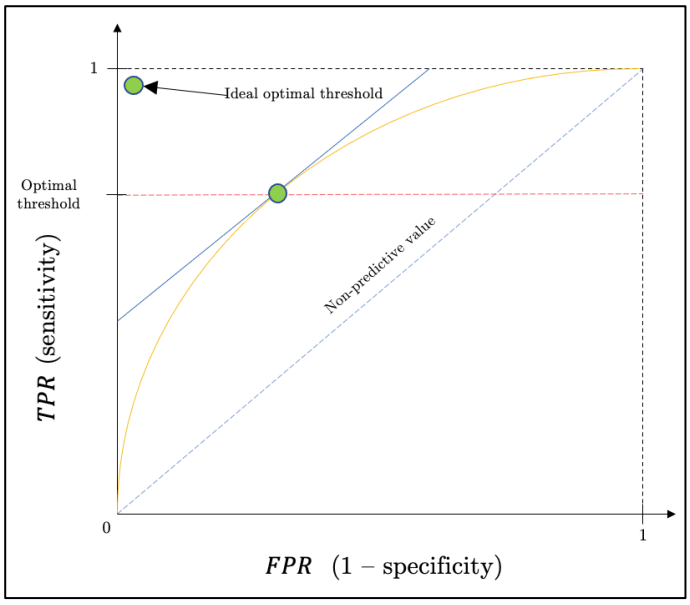

Figure 8 - General setup of a Receiver operating characteristics curve

4.6.2 Features and trees

As mentioned, RF has two internal hyperparameters, namely, the number of iterated trees and the size of the set of features to choose from at each split (i.e., B and m respectively). Firstly, B is set according to the literature of (Oshiro, Perez, & Baranauskas, 2012). The authors argue that as the number of trees grows, the performance does not necessarily increase; beyond a certain point, it only increases computational costs. Oshiro et al. (2012) calculated the Area Under the Curve (AUC), which is a ROC curve-based measure to compare the performance of models, for 29 different data sets. For every data set, the authors compared the AUC of RFs where the number of iterated

trees B is 2^x. In here, x is 1 up to and including 12. The results show that the performance of the models improves up to 128 trees, but after that, the performance does not significantly improve. Therefore, in this research, B is set at 128 iterated trees.

Secondly, based on the lowest OOB estimate of error rate, m is determined. This implies that for every chosen m, the OOB error rate is calculated, thereby taking into account that m cannot exceed the number of features in the data set and a rule-of-thumb is to choose $m \sim \sqrt{p}$.

4.7 Model performance

To assess and compare the performance of the binomial regression and the RF, three statistics are consulted. In here, performance is defined as the ability of a model to predict correct outcomes. The first statistic that is consulted is accuracy and can be seen in formula 16. As mentioned, accuracy is the percentage of outcomes that is correctly predicted. In addition to that, the sensitivity and specificity of a model are consulted (formula 14 and 15). Since the goal of this research is to prove that peer effects exist in certain observations, predicting values of 'one' (i.e., indicating that peer effects exist in an observation) is crucial. This implicates that a model which performs better in predicting 'ones' than in 'zeros', is favoured. Sensitivity is also regarded as the percentage of correctly predicted 'ones' among all predicted 'ones'.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

The accuracy, sensitivity and specificity of the models are measured using a test data set. In doing so, the models are less biased towards training data and the performance is assessed on OOB data. As mentioned before, RF does not require test data to assess the performance of a model. This is because each bagged tree in an RF is trained on approximately 63% of the data. After bagging all trees, the performance is assessed using the remaining 37% of the observations. This is called the OOB estimate of error and can be used as a statistic the assess its performance differences between RFs (Rosenberg, 2017). For comparing the performance of the RF and the binomial regression model, accuracy is used as statistical measure. In this manner, an unbiased comparison method of both models is established.

5. Results

In the following section, the results will be discussed. Firstly, the results of the linear regression model are given. These results will reveal the relationship between the DV, main effect, and moderators. This will provide the basis for answering SRQ1, SRQ2 and SRQ3. After that, the results of the binomial regression model and RF model are presented. For both models, the hyperparameters are optimized to achieve the highest possible performance. Lastly, to determine which model performs better, the accuracy, sensitivity and specificity of both models are compared.

5.1 Linear regression model

The objective of the linear model (table 14 in appendix D) is to explain the dependent variable through 21 independent variables. The dependent variable is *playtime increase player* and the independent variables are separated into relationship characteristics and game characteristics. The former consists exclusively of continuous variables and the latter consists of continuous and discrete variables. Because this research is focused on finding the moderator terms - that affect the relationship between *playtime of players* and *playtime of peers*, the estimates of the interaction terms between main effect (i.e., *playtime increase peer*) and the independent variables are of great importance. In the remainder of this section, the estimates of the relationship characteristics and game characteristics will be discussed. These values provide the basis for answering SRQ1, SRQ2 and SRQ3.

5.1.1 Relationship characteristics

The relationship characteristics consist of the variables: *playtime increase peer, number of peers* and *Euclidean distance.* The estimates of the linear regression show that an increase in these variables increases the DV. This implicates that if *playtime increase peer, number of peers* or the *Euclidean distance* increase by one, *playtime increase player* increases by 0.61, 0,22 or 0.01, respectively and ceteris paribus (table 14, appendix D). Thereby taking into account that the actual increase in the playtime of players is different because all variables are on a logarithmic scale. Besides that, all three variables are significant on the 99%-confidence interval.

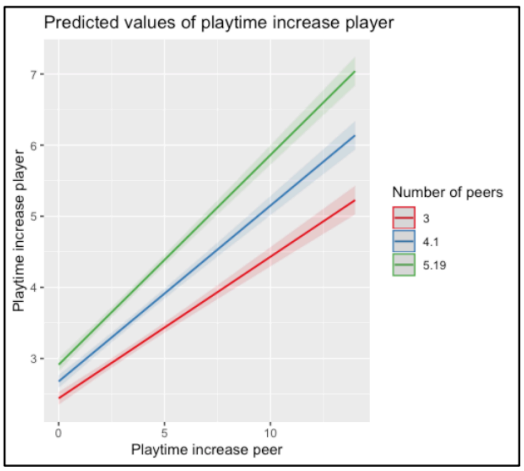

Figure 9 - Interaction plot of number of peers

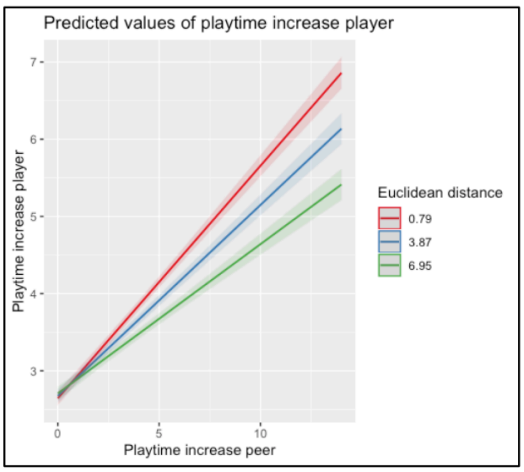

Figure 10 - Interaction plot of Euclidean distance

It is, however, more important to review the estimates of the relationship characteristics that interact with *playtime increase peer*. These interaction terms represent the moderator effect of *number of peers* and *Euclidean distance* on the relationship between *playtime increase player* and *playtime increase peer* (i.e., peer influence). Therefore, the interaction terms *playtime increase peer * number of peers* and *playtime increase peer * Euclidean distance* are included in the linear regression. The estimate of the first interaction term *playtime increase peer * number of peers* is 0.04 (table 14, appendix D). This implicates that an increase of *number of peers* by one, ceteris paribus, results in an increase of the effect of *playtime increase peer* on the DV. The second interaction term *playtime increase peer * Euclidean distance* has an estimate of -0.018. This indicates that an increase of the *Euclidean distance* by one, ceteris paribus, results in a decrease in the effect of *playtime increase peer* on the DV. In figure 9 and 10, the effect of the interaction terms *number of peers* and *Euclidean distance* can be seen. The figures show the relation between *playtime increase player* and *playtime increase peer* for the highest value, lowest value and mean of the *number of peers* and *Euclidean distance*. Besides that, the 95% confidence interval is represented by the shaded region and the slope of the curve and peer effects have a positive relation, meaning that a steeper curve is an indicator of stronger peer effects. In figure 9, it can be seen that an increase in the *number of peers* also increases the slope of the curve. In contrast to that, figure 10 shows the exact opposite relation. Namely, if the *Euclidean distance* decreases, the slope of the curve increases.

5.1.2 Game characteristics

The game characteristics are divided into three categories: general variables, *online capabilities* and *genres*. First of all, the general variables consist of *price, recommendations, achievements* and *DLC* and have estimates of -0.07, 0.40, 0.08 and 0.22 respectively. This implicates that price has a negative relation with the DV and the other variables have a positive relationship with the DV. For a clearer estimation of the interaction effects, the interaction plots can be seen in figures 11 – 14. Figure 11 shows that the slope of the curve increases when the price of a game decreases. This implicates that players are more inclined to follow a peer's behaviour when the price of a game is lower. Besides that, figure 12 demonstrates that the slope of the curve increases when a game has fewer recommendations, which implies that players are more inclined to follow each other's behaviour when a game has fewer recommendations. This assumption also holds for the number of achievements. Namely, the slope of the curve increases when the number of achievements of a game decrease (figure 13). In contrast to that, the number of DLCs and peer effects have a positive relationship; the slope of the curve increases when the number DLCs also increase. This is, however, a limited effect since the differences in slopes are little.

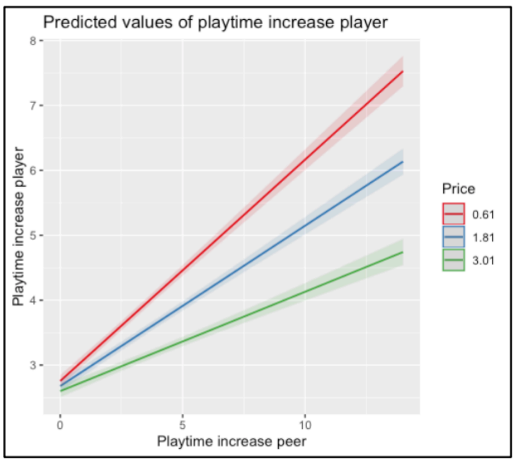

Figure 11 - Interaction plot of price

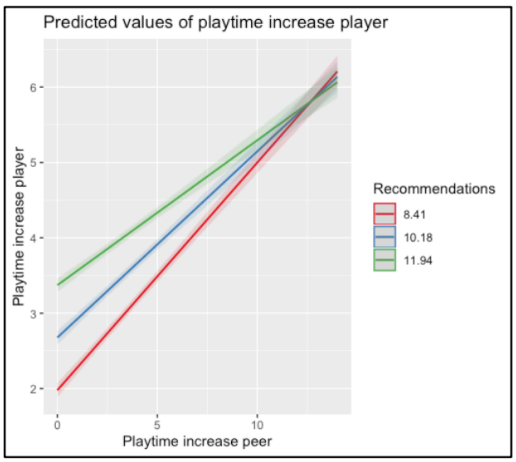

Figure 12 - Interaction plot of recommendations

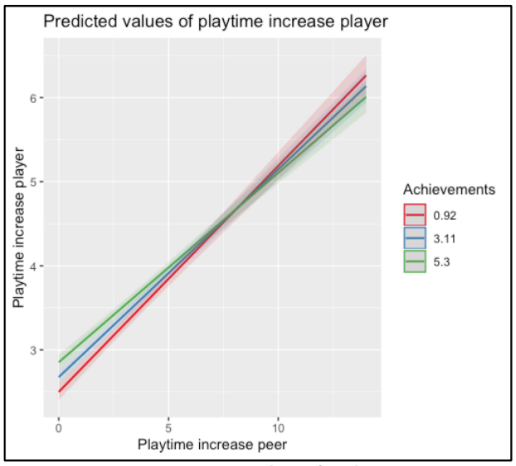

Figure 13 - Interaction plot of achievements

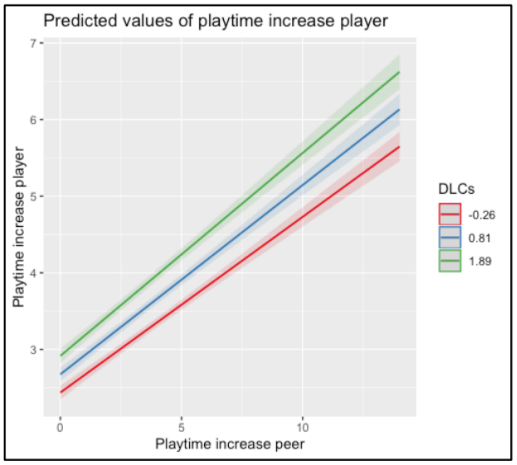

Figure 14 – Interaction of plot DLCs

Second of all, the online capabilities include seven binary variables, namely, *multiplayer, PvP, co-op, online co-op, cross-platform multiplayer, LAN PvP* and *shared/split-screen.* All individual estimates show a positive relationship with the DV, except for *online co-op,* which has a negative relation with the DV (table 14, appendix D). In figures 15 - 20, the plots of the interaction terms, that are significant, can be seen. Figure 15 demonstrates that *multiplayer* games significantly increase the slope. This implicates that peer effects are stronger in games that offer a multiplayer experience. Furthermore, figure 16 shows two interaction plots that are nearly parallel. This implicates that the effects of *PvP* in a game are neglectable. The same assumption can be made for games with *co-op* capabilities (figure 17); this plot also shows two lines that are nearly parallel. In contrast to that, the effects of *Cross Platform Multiplayer* games are more noticeable (figure 18). However, the possibility to play 'cross platform' has a counterintuitive effect. Namely, this online capability decreases peer effects. Moreover, if a game offers *LAN PvP*, *increase playtime players* and *playtime increase peer* show an opposite relation. This implies that for every additional minute a player invests in a game, a peer does not (figure 19). Lastly, figure 20 shows that peer effects are less present when a game offers the possibility to play *shared/split-screen.* It is worth mentioning, however, that a game can belong to more than one genre as well as to none of the genres. If the latter case occurs, all binary variables of the genres are zero and the game belongs to another genre than included in the regression analysis. These observations are captured by the intercept.

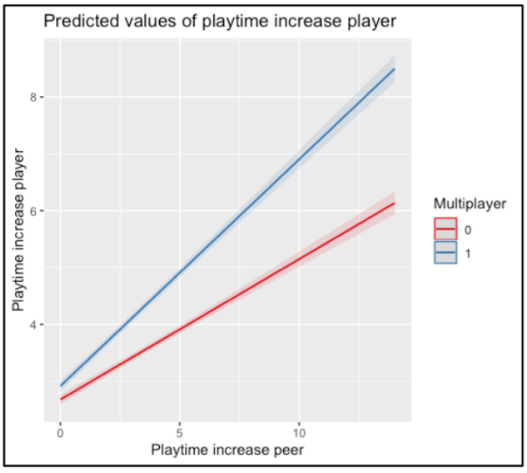

Figure 15 - Interaction plot of multiplayer

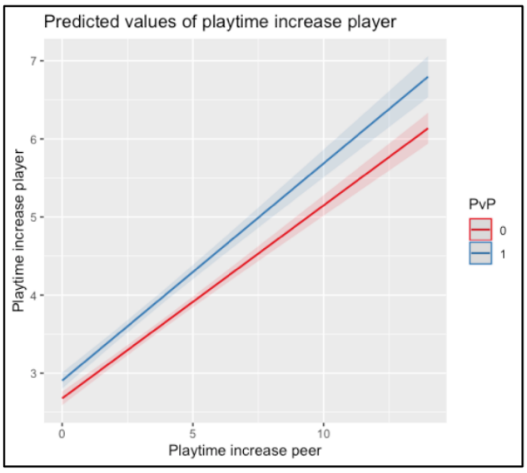

Figure 16 - Interaction plot of PvP

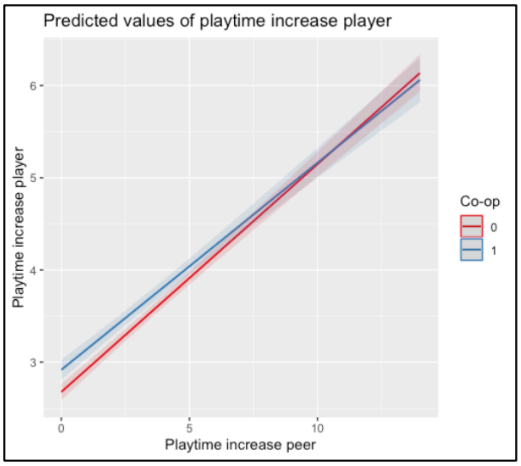

Figure 20 - Interaction plot of co-op

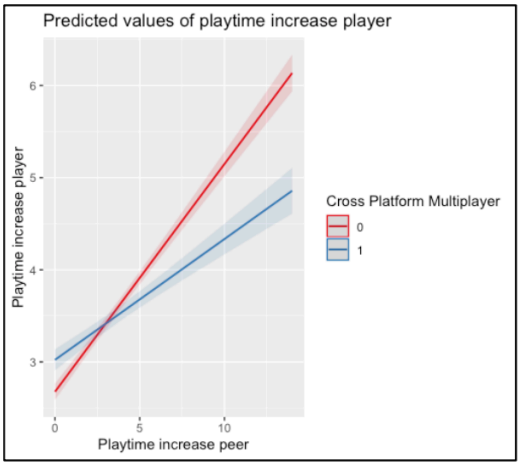

Figure 18 - Interaction plot of Cross Platform Multiplayer

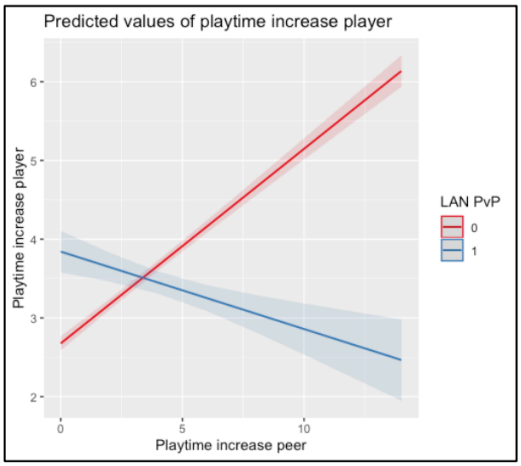

Figure 17 - Interaction plot of LAN PvP

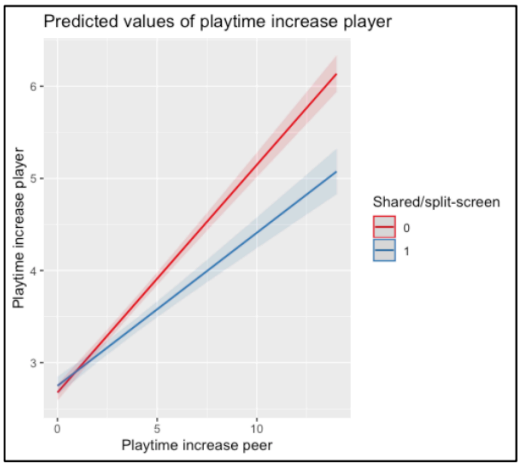

Figure 19 - Interaction plot of shared/split-screen

Lastly, a game can belong to one or multiple genres. The linear regression includes seven binary variables that indicate to what genre(s) a game belongs. The genres are *action, adventure, indie, RPG, free-to-play, strategy* and *simulation.* The estimates of the variables *action* and *indie* have a negative relation with the DV, and the remaining variables have a positive relationship with the DV (table 14; appendix D). All of the above-mentioned variables are significant except for *simulation.* In figures 21 - 25, the interaction plots can be seen. Figures 21 and 22 indicate that peer effects in *adventure* and *strategy* games are less present than in games that are not *strategy* or *indie.* In contrast to that, figures 23 and 24 demonstrate that peer effects are more present in *indie* and *RPG* games when compared to games that are not. However, all four figures (figures 21 - 24) reveal that the slopes of the curves are less responsive to the interaction terms. Hence, the interaction effects of *adventure, indie, RPG* and *strategy* on the relation between *playtime increase player* and *playtime increase peer* are little. Only in figure 25, a major interaction effect can be seen. Namely, peer effects are less present in *free-to-play* games. This implicates that the relation

between *playtime increase player* and *playtime increase peer* is more significant in paid games. Similar to the online capabilities, a game may belong to another genre than the ones included in the linear regression. These observations are captured by the intercept of the regression.

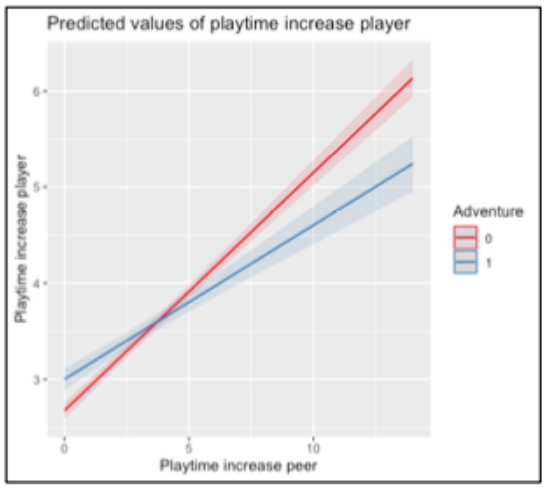

Figure 21 - Interaction plot of adventure

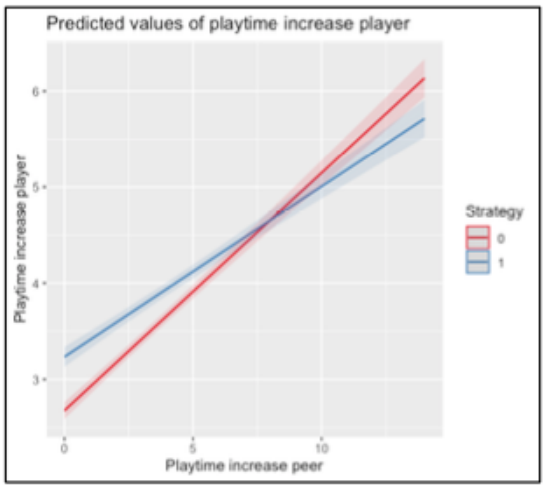

Figure 22 - Interaction plot of strategy

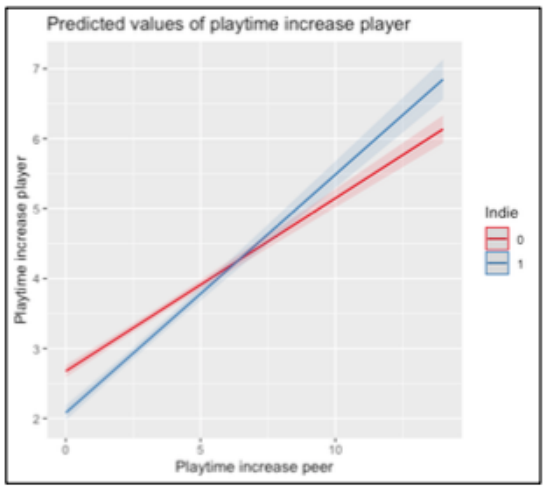

Figure 23 - Interaction plot of indie

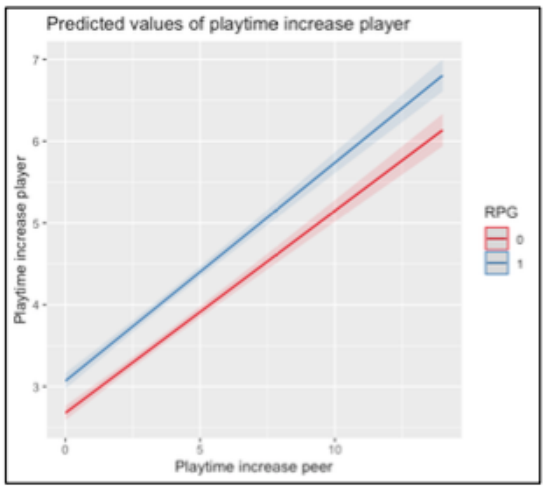

Figure 24 - Interaction plot of RPG

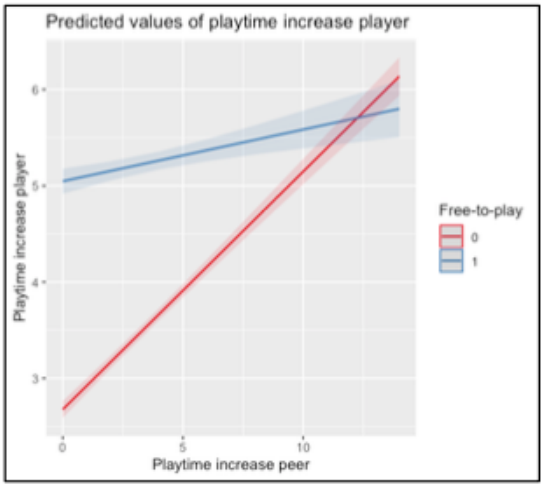

Figure 25 - Interaction plot of free-to-play

5.2 Binomial regression

In the following section, the results of the binomial regression are discussed. The main purpose of this model is to achieve high accuracy. In order to succeed, several statistical methods are used. Firstly, two modifications are performed on the data, namely, an alteration of the DV and a redistribution of positives and negatives. After that, the quality of the model is assessed through information criteria and, lastly, the optimal threshold of the predicted values is determined through the ROC curve.

5.2.1 Rebalancing dependent variable

The binomial regression model is a binary regression model with $n = 1$ Bernoulli trials. Such a model requires a binary dependent variable. Therefore, the continuous variable that is used in the previous section as DV (i.e., *playtime increase player),* is transformed into a binary counterpart. This variable is called *playtime binary* and assumes the value 1 if *playtime increase player* is higher than zero. In table 7, the distribution of zeros and ones can be seen. This shows that the distribution of the variable is unbalanced. Namely, the variable has significantly more negative observations than positive observations and this might result in a model that is severely biased towards negative predictions. To prevent this from occurring, ROSE is applied to the data set. This method combines oversampling and undersampling through bootstrap re-sampling. It increased the number of positive observations and decreased the number of negative observations. As a result, the data set has roughly the same number of positive observations as negative observations.

	Count	*Percentage*
Playtime binary	116,154	44.61
Playtime binary (ROSE)	130,129	49.97

Table 7 - Implications of ROSE redistribution

5.2.2 Information Criteria

After transforming the DV, several methods are used to optimize the model. First, Akaike's Information Criteria and Bayesian's Information Criteria are used to assess the quality of the model. These methods improve the quality of the model by excluding variables. Because BIC assigns a higher penalty to more complex models, this metric favour smaller models over larger models. This implicates that, according to BIC, more variables should be left out of the model compared to AIC. In table 8, the variables that are excluded from the model can be seen. Both criteria determine that

in the optimal model, the variable *playtime increase peer * action* should be excluded from the model. However, BIC proves that also the variables *playtime increase peer * number of peers; playtime increase peer * achievements; playtime increase peer * PvP; playtime increase peer * online co-op; playtime increase peer * RPG* and *playtime increase peer * simulation* should be excluded. In table shows which variables are excluded based on each information criteria. Because the binary regression model does not serve as an explanatory model, less priority is given to the interpretability of the model. Therefore, the results of BIC will be used as reference point for determining the optimal set of predictors.

Akaike's IC	*Bayesian's IC*
Playtime increase peer * action	Playtime increase peer * action
	Playtime increase peer * number of peers
	Playtime increase peer * achievements
	Playtime increase peer * PvP
	Playtime increase peer * online co-op
	Playtime increase peer * RPG
	Playtime increase peer * simulation

Table 8 – Removed variables based on different Information Criteria

5.2.3 ROC curve

Secondly, the data set is split into a training and test set. Seventy per cent of the observations is assigned to the training set and thirty per cent is assigned to the test set. Next, the threshold of classification that optimizes the trade-off between sensitivity and specificity, is determined through the ROC curve. This implies that any threshold, other than the optimal, would deteriorate the sensitivity more than it would improve the specificity and vice versa. Through the ROC curve, it is determined that the optimal threshold is 0.49 (table 15; appendix E). At this value, the sensitivity of the model is 0.79, which implicates that 79 per cent of the predicted values is a correctly predicted positive value. This is also known as the True Positive Rate. Besides that, the specificity of the model is 0.75, implicating that 75 per cent of the predicted values is a correctly predicted negative value. This is also known as the True Negative Rate. The full representation of the ROC curve can be seen in appendix E and the full performance of the model will be assessed in section 5.6.

5.3 Random Forest

Throughout literature, a Random Forest model is regarded as a model with high predicting performance. One of the characteristics of this model is that it requires a binary response variable.

Hence, the binary counterpart of *playtime increase peer, playtime binary,* is used as DV. A downside of this model is that its high performance comes at the expense of interpretability. This model aims to achieve the highest accuracy, specificity and sensitivity. To achieve this, two internal parameters are tuned. In the remainder of this section, tuning of the internal parameters is discussed. In section 5.6, the actual performance of the model will be assessed.

5.3.1 Tuning parameters

In order to optimize the outcome of the RF, two transformations are performed on the data set and two hyperparameters are tuned. First of all, the variable *playtime binary* – which is also used in the binomial regression model - is used as binary DV. Besides that, the data set is split into a training and test set. This split allocates 70 per cent of the observations to the training set and 30 per cent of the observations to the test set. Normally, it is not required to split the data set in the process of training an RF. However, this research will compare the predictive capabilities of the RF with those of the binomial regression. This comparison is most reliable if both models are trained on the same data set. In contrast to that, the predictive capabilities of different RFs are compared by means of OOB error.

After the transformations, two hyperparameters are tuned. Namely, the number of trees and the number of features at each split. The first hyperparameter that is tuned, is the number of trees. Initially, the RF contained 1000 trees and 2 features at each split. This model had an OOB error of 20.60 and is used as a reference point for the future tuned RFs. This implies that 20.60 per cent of the predicted outcomes are incorrect when OOB data is used as input. Based on literature, the number of trees is decreased to 128. This resulted in a significantly decreased computational expense which is highly beneficial for tuning the number of features. It, however, comes at the expense of a slightly increased OOB error. Namely, the error increased with 0.04 per cent (i.e., 20.64 per cent), implying that the model predicts several additional incorrect outcomes.

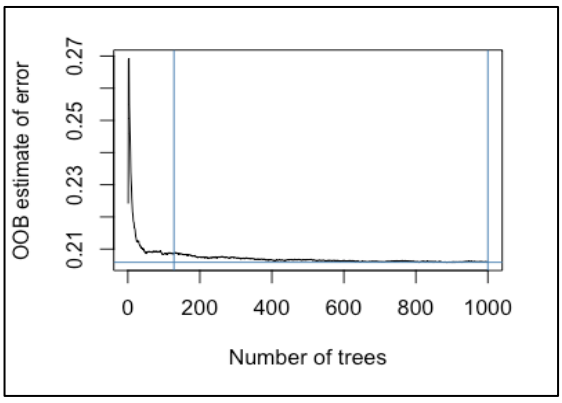

Figure 9 - Optimalisation of number of trees

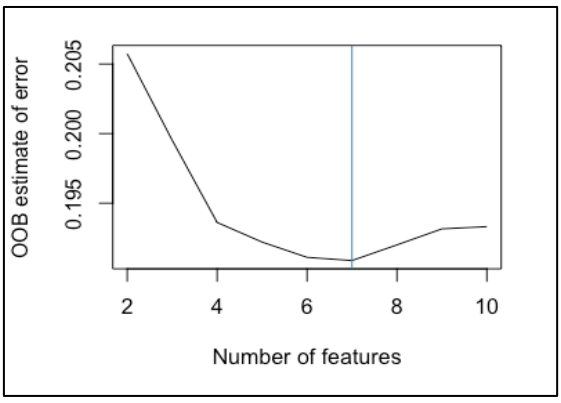

Figure 10 - Optimalisation of number of features

Secondly, the number of features at each split is tuned as a hyperparameter. In order to find the optimal number of features, the OOB estimate of error rate is plotted against the number of features. All RFs have 128 trees; however, they vary in the number of features. In figure 27 it can be seen that 7 features at each split yield the lowest OOB error of 19.24 per cent. This indicates that 19.24 per cent of the predictions are incorrect when OOB observations are used. In table 9, an overview of all RFs can be seen. It shows that the ideal RF has 128 trees and 7 features at each split. This is also the model that will be compared to the binomial regression in section 5.4.

	Random Forest 1	*Random Forest 2*	*Random Forest 3*
Number of trees	1000	128	128
Number of features	2	2	7
OOB error	20.60%	20.64%	19.24%

Table 9 - Comparison of different Random Forest models

5.3.2 Variable importance

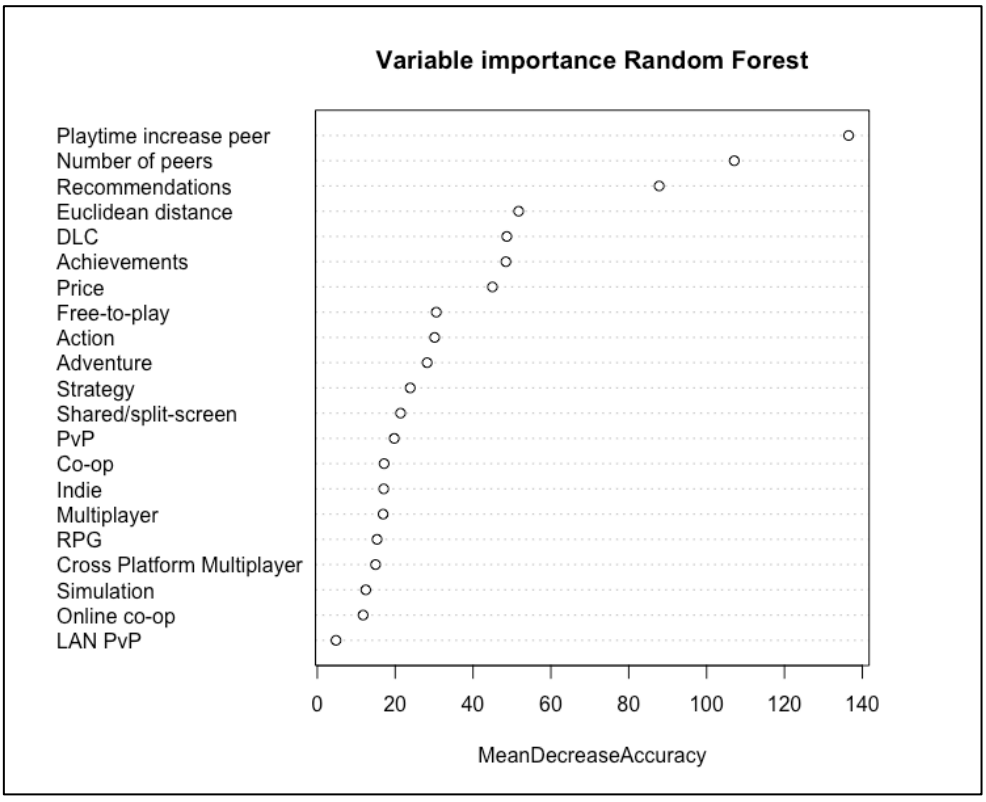

Figure 11 - Variable importance of Random Forest

The OOB error that is discussed in the previous section, is interchangeable with accuracy. Namely, accuracy is the inverse of OOB error. The level of accuracy is established through the predictive capabilities of the variables, and some variables are more important than others. In figure 28, the variable importance is visually represented. This plot shows how the accuracy is affected if

a specific variable is omitted from the RF. As can be seen, *playtime increase peer* affects the accuracy the most, followed by *number of peers, recommendations, Euclidean distance, DLC, achievements* and *price.* Of those variables, *number of peers* and *recommendations* are significantly more important than *Euclidean distance, DLC, achievements* and *price.* Despite that, all continuous variables affect the accuracy to a higher extent than the discrete variables. The variable importance plot shows that *free-to-play, action*, *adventure* and *strategy* are the most important discrete variables. Based on that, it is assumed that genres are more important – in terms of accuracy – than online capabilities.

5.4 Performance comparison

In this section, the performance of the binomial regression model and the Random Forest model is compared. Three different statistics are assessed and based on that, the best model for predicting *playtime binary* is determined. First of all, the Random Forest model has an accuracy of 80.95 per cent, implicating that 80.95 per cent of all predictions are correct (i.e., negative and positive predictions). The accuracy of the binomial regression is 3.98 per cent lower, implying that overall, a binomial regression is worse in predicting *playtime binary.* Secondly, the specificity of the RF is 83.35 per cent, which is 8.00 per cent higher than the specificity of the binomial regression. This implies that the RF is more capable of predicting cases where no peer effects are present. Lastly, the sensitivity of the Random forest is, at 78.35 per cent, 0.64 per cent lower than the sensitivity of the binomial regression. Based on that, it can be determined that a binomial regression is better in identifying cases where peer effects are present. Because the difference in sensitivity is minimal, the differences in accuracies, are mainly caused by the inequality of specificity. However, since this research aims to identify cases where peer effects are present, as well as where they are not present, accuracy is favoured over specificity and sensitivity.

	Binomial regression	*Random Forest*	*Difference*
Accuracy	76.97%	80.95%	3.98%
Specificity	75.35%	83.35%	8.00%
Sensitivity	78.99%	78.35%	-0.64%

Table 10 - Comparison of different predictive models

5.5 Reflection on the hypotheses

Based on the results of the linear regression model, binomial regression model and Random Forest, the hypotheses that are stated in section 2.3 and 2.4, can be reflected upon. In the following section, the findings of gaming behaviour - and its moderators – are given.

5.5.1 Gaming behaviour

To give a more precise answer to the research question, it is divided into four sub-questions. Firstly, *"What is the main effect of gaming behaviour of an individual?"* (i.e., S-RQ1). It is hypothesised that the main effect of an individual's gaming behaviour is the gaming behaviour of one's peers (i.e., H1). The linear regression model shows that *playtime increase player* – which is the dependent variable – rises when *playtime increase peer* increases. Besides that, *playtime increase peer* is the most important variable in the RF. Based on these observations, the hypothesis is accepted.

5.5.2 Relationship characteristics

However, the relation between *playtime increase peer* and the DV is highly moderated by several other predictors, associated with the relationship between peers. To examine this, the following sub-question question is used: "What characteristics of the relationship between peers, moderate gaming behaviour?" (i.e., S-RQ2). Based on this research question, two hypotheses are proposed. The first hypothesis states that the *number of peers* moderates gaming behaviour, such that peer effects are stronger when a gamer has few friends (i.e., $H2_{a)}$. The interaction plot of *number of peers* show that more peers result in a stronger relationship between the gaming behaviour of peers. As a result, this hypothesis is rejected. The second hypothesis states that the similarity between gamers moderates gaming behaviour, such that peer influence is stronger when two gamers are more alike (i.e., $H2_b$). The interaction plot in the previous section shows that a higher *Euclidean distance* results in a stronger relationship between the gaming behaviour of peers. A higher *Euclidean distance* is an indicator for peers that are less alike. Therefore, the hypothesis specific to this research question is accepted.

5.5.3 Game characteristics

Besides relationship characteristics, peer effects are also moderated by the characteristics of a game. Therefore, the following sub-question is used: "What characteristics of a game moderate gaming behaviour?" (i.e., S-RQ3). Based on that, six hypotheses are proposed. First of all, it is assumed that the *price* of a game moderates gaming behaviour in such a manner that peer influence is stronger when a game has a higher price (i.e., $H3_{a)}$. However, the interaction plot of this specific variable shows that peer effects are stronger when the *price* of a game decreases. As a result, this hypothesis is rejected. Second of all, the amount of DLC that is available in a game, is expected to moderate gaming behaviour, such that peer influences are stronger if more DLC is available (i.e., $H3_b$). The interaction plots show that more DLC leads to stronger peer influences, however, these effects are limited. Despite that, the hypothesis is accepted due to the small positive change in peer influence. Furthermore, it is hypothesised that *recommendation* and *achievements* moderate gaming behaviour, such that peer influence is stronger in games with more recommendations and achievements (i.e., $H3_c$ and $H3_{d)}$. Both interaction plots of these variables contradict this assumption, however, these effects are limited as well. Nonetheless, the hypotheses are rejected due to the small negative change in peer influence.

The fifth hypothesis states that *online capabilities* moderate gaming behaviour, such that peer influences are stronger if certain capabilities are available in a game (i.e., $H3_e$). Based on the interaction plots it can be determined that peer influences are stronger in *multiplayer* and *PvP* games. In *co-op, Cross Platform Multiplayer, LAN PvP* and *shared/split-screen* games, peer influences are weaker. Therefore, the hypothesis is partially true and is, hence, accepted. The sixth hypothesis states that the *genre* of a game moderates gaming behaviour, such that peer influences are stronger for certain genres (i.e., $H3_f$). Although *adventure* and *strategy* games have a negative effect, and *RPG* and *indie* have a positive effect on the magnitude of peer influences, the interaction plots demonstrate that the effects of these four genres are limited. Only in *free-to-play* games, the peer influences significantly less. Because most genres have a limited or negative effect on the magnitude of peer effects, the hypothesis is rejected.

5.5.4 Model performance

Lastly, a sub-question concerning the performance of the predictive models is formulated. Namely, "When predicting peer effects, a Random Forest can achieve the highest performance" (i.e., S-RQ4). It is hypothesised that a Random Forest can achieve the highest predicting performance (i.e., H4). Based on the accuracy of the Random Forest and the binomial regression, it can be

determined that an RF is indeed the better performing model. Hence, the hypothesis is accepted. In table 11, an overview of the correctness of the hypotheses can be seen.

Hypothesis	*Accepted/rejected*
H1	Accepted
$H2_a$	Rejected
$H2_b$	Accepted
$H3_a$	Rejected
$H3_b$	Accepted
$H3_c$	Rejected
$H3_d$	Rejected
$H3_e$	Accepted
$H3_f$	Rejected
H4	Accepted

Table 11 - Overview of accepted/rejected hypotheses

5.6 Updated conceptual framework

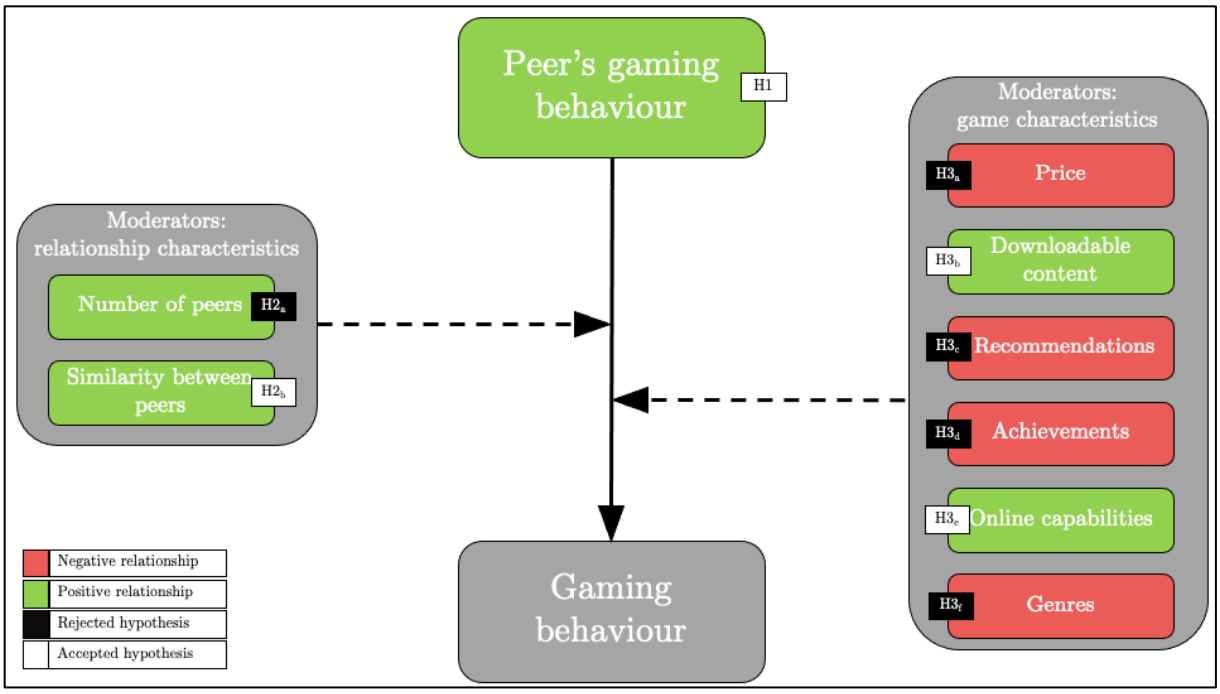

Figure 12 - Updated conceptual framework

Section 5.5 shows that some outcomes were not expected in the first stages of this research. As a result, some hypotheses are rejected. This deemed the initial conceptual framework, as resented in section 2.5, incorrect. Therefore, this conceptual framework is updated. In figure 29, the revised conceptual framework can be seen. It shows that peer's gaming behaviour has a positive effect on gaming behaviour of their peers, but it is moderated by the characteristics of the relationship and the game. In the conceptual framework, these moderator effects are represented by the dotted line.

On the left-hand side, the *number of peers* and similarity between peers (i.e., *Euclidean distance*) is shown. The new framework shows that both characteristics positively moderate the relationship between gaming behaviour of peers. On the right-hand side, it can be seen that *price, recommendations, achievements* and *genres* negatively impact the relationship between gaming behaviour of peers. *DLC* and *online capabilities*, however, have a positive impact on this relationship.

6. Discussion

In the previous sections, it is determined whether the hypotheses are accepted or rejected. This section will discuss if the accepted hypotheses are in accordance with literature. If the hypotheses are rejected, an explanation will be provided. Additionally, the accepted and rejected hypotheses are visually represented in figure 29.

6.1 Accepted hypotheses

Five out of ten hypotheses in this research are accepted. Firstly, the general hypothesis that gaming behaviour of one's peers affects an individual's gaming behaviour, is accepted. Amialchuk and Kotalik (2016) stated that gamers tend to invest additional hours in a game when their peers increase the in-game time of that particular game. The authors argue, however, that the effects are more exogenous when a study uses nominated peers. Therefore, Amialchuk and Kotalik (2016) used grade-level peers. In this research, the exogenous effects of this relationship are represented by the moderator terms. So, even though nominated peer selection is used, a distinct effect of playtime of peers on playtime of players can be observed.

Secondly, the hypothesis that similarity between peers increases the effect of one's gaming behaviour on gaming behaviour of their peers, is accepted. This supports the findings of Wu, Chen and Chung (2010) as they state that shared values of social structures and between peers have a positive effect on the trust in, and relationship with a social structure. Also, the outcomes of Achananuparp, & Lim (2012) are supported as they state that community structures have an increased effect on social contagion. This social contagion is the strength of the peer effect.

Besides that, the hypothesis that the availability of *DLC* positively affects the strength of peer effects is accepted. According to McCaffrey (2019), the addictive properties of *DLCs* might increase in in-game time. In that manner, a peer might influence its peer network into investing more time into playing a certain game. As a result, the relationship between gaming behaviour of two peers is stronger.

Furthermore, the hypothesis that the *online capabilities* of a game positively affect the strength of the relationship between two peers, is accepted. Thereby it must be taken into account that not all *online capabilities* strengthen peer influences. Lemmens and Hendriks (2016), showed that the differences between online and offline – in terms of in-game time – are not significant. However, when comparing an online game with an offline game within one genre, the differences are more significant. The results of the linear regression show that *PvP* and *co-op* have neglectable effects on peer influence. The effects of these variables might be captured by the *multiplayer* variable, which shows to have a significant positive effect on peer influence. Games that offer multiplayer capabilities have increased social elements. This might increase the urge of gamers to interact with each other and, hence increase the in-game time. Interestingly, games that offer *Cross Platform Multiplayer* capabilities are less likely to strengthen peer influences. However, the fact that these games are playable 'cross platform' might implicate that they are played with gamers from outside the gaming platform (i.e., Steam). Since this research only identifies peer effects between gamers on the same platform, these effects are not measured. Similar effects can be observed for *shared/split-screen* games. These games, however, facilitate the ability to play games on one PC. Therefore, the in-game time of one player is measured and peer effects are not identified. The strongest negative effect on peer influences is observed for *LAN PvP*. It is assumed that these effects are measured because these games support gamers to play against each other without being connected to the internet. Namely, they use a Local Area Network (LAN) connection to link their PCs. In this manner, the gaming platform fails to recognize that the in-game time of a gamer increases.

Lastly, the hypothesis that a Random Forest performs better than a binomial regression when predicting peer effects in a gaming environment, is accepted. Throughout literature, this method is highly preferred. In terms of accuracy, it is clear that an RF is the better predicting model. Still, a binomial regression is not a poor alternative since it performs slightly better in terms of sensitivity.

6.2 Rejected hypotheses

The remaining five hypotheses in this research are rejected based on the fact that the models show opposing results. Firstly, the hypothesis that peer effects are stronger when a gamer has few peers, is rejected. Literature suggests that a gamer with few peers has less influential power of its peers (Zsolt, Zubcsek, & Sarvary, 2011). This is, however, not the case. The interaction plot of *number of peers* show that peer effects are, in fact, stronger when a gamer has more peers. Possibly,

gamers with many peers are the opinion leader of a social structure (De Veirman, Hudders, & Nelson, 2009). In that manner, these gamers are more likely to influence other gamers in their immediate environment.

Secondly, it is hypothesised that a higher *price* results in stronger peer effects. Based on the results, this hypothesis is rejected. The interaction plot shows that a lower *price* results in stronger peer effects. Gamers may be more likely to copy a game from their peers when this game has a lower *price*. In that sense, it contradicts the findings of Liao, Tseng, Cheng, & Teng (2020), who argue that perceived price fairness is positively related to gamer loyalty. Interestingly, games that belong to the *free-to-play* genre, seem to weaken peer effects. It is expected that the *free-to-play* games in this research have less appeal to the social aspects of gaming, and hence do not increase peer effects. This is, concurrently, the only genre that has a significant effect on the strength of peer effects, albeit that this effect is negative. As a result, the hypothesis that certain genres have a positive effect on the strength of peer effects is rejected.

Furthermore, the hypotheses that many *recommendations* or many *achievements* result in stronger peer effects, are rejected. Based on the interaction plots, it can be determined that the effects of *recommendations* and *achievements* are minimal. Moreover, these variables affect peer effects in such a manner that more *recommendations* or more *achievements* result in weaker peer effects. Literature suggests that a game with many *recommendations* is bought more often and more easily (Senecal & Nantel, 2004). This assumption might still hold, albeit that it does not significantly change the strength of peer effects. With regard to *achievements*, Hamari (2014) suggests that achievements have a positive effect on the engagement of a gamer towards a game. Equal to the *recommendations* of a game, this assumption might still hold for an individual gamer. However, it does not affect the strength of peer effects.

7. Conclusion

7.1 Research structure

This research started by exploring the current state of gaming environments and aims to identify peer effects in a gaming environment. In order to do so, the following research question is stated: *"What factors contribute to gaming behaviour in Steam's gaming environment?"*. This research question is divided into four sub-questions. Three of those sub-questions focus on the main effect and moderators of gaming behaviour, and one aims to estimate the accuracy of a predictive

model. In section 2, an extensive literature study is presented. Previous research shows that gaming behaviour can be affected by gaming behaviour of one's peers.

However, this behaviour is severely moderated by several characteristics. These characteristics are divided into relationship characteristics and game characteristics. By means of existing literature, two relationship characteristics are identified, namely, the *number of peers* (Katz & Lazarsfeld, 1955) and the similarity between peers (Wu, Chen, & Chung, 2010). Furthermore, several game characteristics are established, namely, *price*, *DLC* (Liao, Tseng, Cheng, & Teng, 2020), *recommendations* (Senecal & Nantel, 2004), *achievements* (Hamari, 2014), *online capabilities* and the *genre* of a game (Lemmens & Hendriks, 2016). Based on existing literature, nine hypotheses are formed which assisted in answering the research questions. In section 2.5, the initial conceptual framework is visualized.

After proposing the hypotheses, the process of data collection is described in section 3. The raw player data originates from the research of O'Neill, Vaziripour, Wu and Zappala (2016). This data was, however, severely coded and needed transformations and replications before it could be used in this research properly. After that, game data is collected from the servers of the Steam Storefront using an API. Both datasets were consolidated into one and, as a result, the constructed data set contained over 2.5 million observations. Through aggregation, a subset of 260,389 observations is created.

Using this data, three statistical models were created. The first model is a linear regression that is used to answer sub-questions one, two and three, whereas the remaining two models are compared with each other. These models form the basis for answering the fourth sub-question.

7.2 Main conclusion

The results of the linear regression model showed that gaming behaviour is affected by the gaming behaviour of its peers. This behaviour, that is also known as peer effects, is severely moderated by characteristics of the relationship between peers and the game. First of all, it is demonstrated that the *number of peers* and the similarity between peers have a positive effect on peer effects. So, as far as the characteristics of the relationship go, a larger friendship network size with similar preferences as the player, will result in gaming behaviour that is more alike. This implicates that it is likely that a gamer will copy the behaviour of its peers.

Furthermore, the linear regression shows a strong effect of a low *price* on gaming behaviour of peers. This implicates that gaming behaviour of peers is more alike when they play an inexpensive game. This is likely because a gamer only has to invest a small amount of money, which makes it

easier for the gamer to copy the behaviour of its peer. In contrast to that, *free-to-play* games have a strong negative effect on the relationship between gaming behaviour of peers. It is expected that the *free-to-play* games in this research, appeal less to the social aspects of gaming, and hence decrease peer effects. Besides the *free-to-play* genre, none of the genres seems to affect the strength of peer effects.

In addition to that, *DLC* has a small positive effect on the strength of peer effects. The linear regression shows that a game with more additional content increases peer effects. This likely is because *DLC* increases the possibilities in a game. Together with the addictive properties of *DLC*, it increases the similarity between gaming behaviour of peers.

The evidence of the linear regression also shows that – with regard to *online capabilities* - *multiplayer* games increase peer effects. Interestingly, games with *LAN PvP* capabilities have an opposite effect on gaming behaviour. This implicates that, if a gamer increases in-game time, its peers decrease their in-game time. It is possible that due to the *LAN* properties of a game, gamers are no longer connected to the internet and hence, their in-game time is not registered. As such, there are no peer effects to observe in such games.

Lastly, the number of *achievements* and *recommendations* of a game have a limited effect on peer effects. It is, however, worth mentioning that the small effect that is observed is negative. It is assumed that *achievements* and *recommendations* only affect the purchase decision of a gamer and not the strength of the relationship between gaming behaviour of peers.

Based on all the remarks in this section, the main research question can be answered. Gaming behaviour is first and foremost affected by gaming behaviour of its peers. This is the main effect of gaming behaviour and is referred to as peer effects. However, gaming behaviour is moderated by various characteristics of the relationship and game. It seems that the number of peers and similarity between peers are important characteristics of the relationship, whereas price, DLC and online capabilities seem to be important factors of the characteristics of the game. These assumptions are captured by the updated conceptual framework in section 5.5. In addition to that, it is determined that a Random Forest is the most capable of predicting peer effects in a gaming environment.

7.3 Limitations

Several limitations are identified in this research. Firstly, the lack of adequate computational power had a significant effect on the computing time. The effects of this power deficit were especially noticed during the process of data collection. The raw player data of this research is previously used

in the research of O'Neill, Vaziripour, Wu and Zappala (2016). Their research had a time span of over a year, as well as the availability of high computational power. Therefore, their research included data of over 18 million players. Because these amounts of data require substantial system memory, it was not possible to use all player data, hence the significant decrease in sample size. Additionally, their data was highly coded. Many transformations and replications were required in order for the data to be used properly. The lack of computational power also decreased the ability to tune the hyperparameters even further. This, however, might not significantly increase the predictive accuracy of the models.

Secondly, this study used nominated peers instead of appointed peers. The differences between these methods are discussed in section 2.1.1, and it can be concluded that both methods have their pros and cons. However, the major downside of nominated peers, compared to appointed peers, is that the correlated effects between peers cannot be separated from the endogenous effects. This might implicate that peer effects occur because one peer chooses another peer, and not due to the characteristics of the game or relationship. Nevertheless, a nominated peer selection method is a commonly used method in scientific research.

A further limitation of this research is that peer effects might be a fixed effect. In that sense, the characteristics or relationship of a game do not affect peer effects; some players are inherently more inclined to affect a peer's behaviour. It is not possible to control for these fixed effects due to the way the data is constructed. Besides that, this research focussed solely on peer effects in Steam's gaming environment. All observations in this study are from the Steam database. Besides Steam, several different gaming environments exist. In doing so, the external validity of this research is compromised, implicating that the remarks of this study might not apply to different gaming environments.

In addition to that, the data that was used for this research dates back from 2014. This might result in a decrease in scientific relevance, albeit that relationship and game characteristics did not significantly change over the years. Lastly, this research was affected by the limited previous research into peer effects in a gaming environment. Even though a lot of research is devoted to identifying peer effects in many different environments, as well as the implications of Internet Gaming Disorder (IGD), a gap exists between the two. In that sense, the specifics of peer effects in a gaming environment remain unclear.

7.4 Directions for further research

As mentioned in section 7.3, a gap exists between literature on peer effects and behaviour in a gaming environment. Even though this research is a first attempt to filling this gap, there is a great deal to discover in this field of study. First of all, more gaming environments such as PlayStation Network, Xbox Live or Nintendo Network could be included. This will increase the external validity of future research. Secondly, the lack of demographic variables should be addressed in future research. Including such variables might be beneficial for future research as it enhances the background characteristics of the target audience. To accomplish this, future research has to overcome several challenges with regard to privacy. The data in this study dates back to 2014. Back then, the limitation due to privacy regulations was less than nowadays. This means that replicating this study with more recent data, will be more challenging. With respect to collecting demographic data, even more challenges must be overcome. Nevertheless, future research might benefit from data that is more recent. In addition to that, data that has more different data collections could increase the robustness of future research. In doing so, it can observe differences in gaming behaviour through time.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov, & F. Csaki, *Second international symposium on information theory* (pp. 267-281). Budapest, Hungary: Akademai Kiado.

Akaike, H. (1976). Canonical Correlation Analysis of Time Series and the Use of an Information Criterion. *Mathematics in Science and Engineering,126*, 27-96.

Amialchuk, A., & Kotalik, A. (2016). Do Your School Mates Influence How Long You Game? Evidence from the U.S. *PloS one 11(8)*.

Argyle, M. (2001). The psychology of happiness. London: Routledge.

Atari. (n.d.). *About us*. Retrieved from Atari: https://www.atari.com/about-us/

Berndt, T. (2002). Friendship Quality and Social Development. *Current Directions in Psychological Science, 11(1)*, 7-10.

Breiman, L. (1996). Bagging predictors. *Machine Learining 26(2)*, 123-140.

Breiman, L. (2001). Random Forests. *Machine Learning 45*, 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* Wadsworth: Belmont.

Ceriani, L., & Verme, P. (2012). The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality, 10*, 421-443.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *Proceedings of the Fourth International AAAI Conference on Weblogs and Socia Media.*

Chikhani, R. (2015, October 31). *The Histrory of Gaming: An Evolving Community*. Retrieved from Techcrunch.com: https://techcrunch.com/2015/10/31/the-history-of-gaming-an-evolving-community/

Clement, J. (2020, October). *Global digital population as of October 2020*. Retrieved from Statista: https://www.statista.com/statistics/617136/digital-population-worldwide/

Couronné, R., Probst, P., & Boulesteix, A. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Biomedics, 19(1)*.

De Veirman, M., Hudders, L., & Nelson, M. (2009). What Is Influencer Marketing and How Does It Target Children? A Review and Direction for Future Research. *Front Psychology, 10(1)*.

Derksen, S., & Keselman, H. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology,45*, 265-282.

Dziak, J., Coffman, D., Lanza, S., & Li, R. (2012). *Sensitivity and specificity of information criteria.* State College, PA: The Methodology Center, The Pennsylvania State University.

Eklund, L. (2012). *The Sociality of Gaming: A mixed methods approach to understanding digital gaming as a social leisure activity.* Stockholm: PhD dissertation.

Elmore, K., & Richman, M. (2001). Euclidean Distance as a Similarity Metric for Principal Component Analysis. *Monthly Weather Review 129*, 504 - 549.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27(8)*, 861-874.

Gini, C. (1912). Variabilità e Mutuabilità. *Reprinted in Memorie di metodologica statistica.*

Gourdon, X., & Sebah, P. (2001, January 2001). *Binary Splitting Method.* Retrieved from Mathematical constants and computation: http://numbers.computation.free.fr/Constants/Algorithms/splitting.ps

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78(1)*, 1360-1380.

Gummerus, J., Liljander, V., Pura, M., & Van Riel, A. (2004). Customer loyalty to content-based web sites: the case of an online health-care service. *J. Serv. Mark., 18(3)*, 175-186.

Hamari, J. (2014). Does Gamification Work? -- A Literature Review of Empirical Studies on Gamification. 47th Hawaii International Conference on System Sciences.

Hamari, J. (2017). Do badges increase user activity? A field experiment on the effects of gamification. *Comuters in Human Behavior, 71*, 469-478.

Hamari, J., & Eranti, V. (2011). Framework for Designing and Evaluating Game Achievements. DiGRA 2011 Conference: Think Design Play.

Harrigan, N., Achananuparp, P., & Lim, E. (2012). Influentials, novelty, and social contagion: The viral power of average friends, close communities, and old news. *Social Networks, 34(4)*, 470-480.

History.com. (2019, June 10). *Video Game History*. Retrieved from History.com: https://www.history.com/topics/inventions/history-of-video-games

Ho, T. (1995). Random Decision Forests. *3rd International Conference on Document Analysis and Recognition* (pp. 278-282). Montreal, QC: AT&T Bell Laboratories.

Huotari, K., & Hamari, J. (2011). "Gamification" from the perspective of service marketing . Gamification Workshop, CHI2011.

Katz, E., & Lazarsfeld, P. (1955). Personal Influence; the Part Played by People in the Flow of Mass Communications. Glencoe, IL: Free Press.

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review, 3(1).*

Kuha, J. (2004). Comparisons of Assumptions and Performance. *Sociological Methods & Research, 33(2)*, 188-229.

Lee, J., Jett, J., & Perti, A. (2015). The Problem of "Additional Content" in Video Games. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries.* New York, NY: Association for Computing Machinery.

Lemmens, J., & Hendriks, S. (2016). Addictive Online Games: Examining the Relationship Between Game Genres and Internet Gaming Disorder. *Cyberpsychology, Behavior, and Social Networking, 19(4).*

Liao, G., Tseng, F., Cheng, T., & Teng, C. (2020). Impact of gaming habits on motivation to attain gaming goals, perceived price fairness, and online gamer loyalty: Perspective of consistency principle. *Telematics and Infomatics, 49(1).*

Malone, T. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 5(4)*, 333-360.

McCaffrey, M. (2019). The macro problem of microtransactions: The self-regulatory challenges of video game loot boxes. *Business Horizons, 62(4)*, 483-495.

Medium.com. (2017, January 20). *The History of Online Gaming*. Retrieved from Datapath.io: https://medium.com/@datapath_io/the-history-of-online-gaming-2e70d51ab437

Menardi, G., & Torelli, N. (2012). Training and assessing classification rules with unbalanced data. *Data Mining and Knowledge Discovery.*

Moore, D., McCabe, G., Alwan, L., Craig, B., & Duckworth, W. (2011). Logisitic regression. In D. Moore, G. McCabe, L. Alwan, B. Craig, & W. Duckworth, *The practice of Statistics for Business and Economics* (pp. 620-621). New York, NY: W.H. Freeman.

Moore, D., McCabe, G., Alwan, L., Craig, B., & Duckworth, W. (2011). Multiple regression. In D. Moore, G. McCabe, L. Alwan, B. Craig, & W. Duckworth, *The Practice of Statistics for Business and Economics* (pp. 581-584). New York, NY: W.H. Freeman and Company.

Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016). Comparing Random Forest with Logistic Regression fo Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis, 24(1)*, 87-103.

NewZoo. (n.d.a). *Top 25 Public Companies by Game Revenues.* Retrieved from NewZoo.com: https://newzoo.com/insights/rankings/top-25-companies-game-revenues/

Nunes, J., & Dréze, X. (2006). Your Loyalty Program Is Betraying You. *Harvard Business Review, 84(4)*, 124-131.

O'Neill, M., Vaziripour, E., Wu, J., & Zappala, D. (2016). *Condensing Steam: Distilling the Diversity of Gamer Behavior.* Brigham: Brigham Young University.

Oshiro, T., Perez, P., & Baranauskas, J. (2012). How Many Trees in a Random Forest? *Lecture notes in computer science.*

Panigrahi, S. (2020). *How to Differentiate Between Peer Pressure and Peer Influence?* Retrieved from Higher Education Review: https://www.thehighereducationreview.com/news/how-to-differentiate-between-peer-pressure-and-peer-influence-nid-1172.html

Rosenberg, D. (2017). *Bagging and Random Forests.* Retrieved from https://davidrosenberg.github.io/mlcourse/Archive/2017/Lectures/9a.bagging-random-forests.pdf

Ryan, C. (2017). Measurement of Peer Effects. *The Australian Economic Review, 50(1)*, 121-129.

Senecal, S., & Nantel, J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of Retailing, 80(2)*, 159-169.

Smithsonian. (n.d.). *The Father of the Video Game: The Ralph Baer Prototypes and Electronic Games.* Retrieved from Smithsonian: https://www.si.edu/spotlight/the-father-of-the-video-game-the-ralph-baer-prototypes-and-electronic-games/video-game-history

Statista. (n.d.a). *Number of active video gamers worldwide from 2014 to 2021.* Retrieved from Statista.com: https://www.statista.com/statistics/748044/number-video-gamers-world/

Statista. (n.d.b). *Gaming revenue worldwide 2020, by segment.* Retrieved from Statista.com: https://www.statista.com/statistics/278181/video-games-revenue-worldwide-from-2012-to-2015-by-source/

Statista. (n.d.c). *Genre breakdown of video game sales in the United States in 2018.* Retrieved from Statista.com: https://www.statista.com/statistics/189592/breakdown-of-us-video-game-sales-2009-by-genre/

Steam. (n.d.a). *Store.* Retrieved from Steam: https://store.steampowered.com

Steam. (n.d.b). *About Steam.* Retrieved from Steampowered.com: https://store.steampowered.com/about/

Steam. (n.d.c). *Steamworks-documentatie.* Retrieved from Steamworks.com: https://partner.steamgames.com/doc/store/application

Steam. (n.d.d). *Steam Web API.* Retrieved from Valve Developer Community: https://developer.valvesoftware.com/wiki/Steam_Web_API#GetNewsForApp_.28v0001.29

Storm van Leeuwen, Q. (2019, 9 30). *Het belang van online reviews: de invloed van klantbeoordelingen op het aankoopgedrag van de consument.* Retrieved from Capterra: https://www.capterra.nl/blog/951/belang-van-online-klantbeoordelingen-op-aankoopgedrag-van-consument

TechTerms. (n.d.a). *DLC Definition.* Retrieved from Techterms.com: https://techterms.com/definition/dlc

The Entertainment Software Association (ESA). (2018). *2018 Sales, Demographics, and Usage Data: Essential Facts About The Computer And Video Game Industry.* Washington, DC: Entertainment Software Association.

The Entertainment Software Association (ESA). (2020). *2020 Essential Facts About the Video Game Industry.* Washinton, DC: ESA.

Uspensky, J. (1937). Bernoulli's Theorem. In J. Uspensky, *Introduction To Mathematical Probability* (pp. 96-101). New York and London: McGraw-Hill Book Company Inc.

Valve Corporation. (n.d.a). *About Steam.* Retrieved from Steam: https://store.steampowered.com/about/

Valvi, A., & West, D. (2013). E-loyalty is not all about trust, price also matters: extending expectation-confirmation theory in bookselling websites. *J. Electron. Commer. Res., 14(1)*, 99-123.

Watts, D., & Dodds, P. (2007). Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research 34(4)*, 441-458.

Webb, G. (2011). Overfitting. In C. Sammut, & G. Webb, *Encyclopedia of Machine Learning.* Boston, MA: Springer.

Williams, M. (2017, October 11). *The Harsh History Of Gaming Microtransactions: From Horse Armor to Loot Boxes.* Retrieved from Usgamer.net: https://www.usgamer.net/articles/the-history-of-gaming-microtransactions-from-horse-armor-to-loot-boxes

World Health Organization. (2018). *International classification of diseases for mortality and morbidity statistics (11th revision).* Retrieved from https://icd.who.int/browse11/l-m/en

Wu, F., & Huberman, B. (2007). Novelty and collective attention. *PNAS, 104(1)*, 17599-17601.

Wu, J., Chen, Y., & Chung, Y. (2010). Trust factors influencing virtual community members: A study of transaction communities. *Journal of Business Research, 63(10)*, 1025-1032.

Xia, L., Monroe, K., & Cox, L. (2004). The price is unfair! A conceptual framework of price fairness perceptions. *J. Mark, 68(4)*, 1-15.

Zagal, J., Björk, S., & Lewis, C. (2013). Dark Patterns in the Design of Games . *Foundations of Digtal Games 2013.*

Zsolt, K., Zubcsek, P., & Sarvary, M. (2011). Network Effects and Personal Influences: The Diffusion of an Online Social Network. *Journal of Marketing Research, 48(3)*, 425-443.

Appendices

Appendix A

Dependent variables		*Independent variables*			
		Game characteristics		*Relationship characteristics*	
Variable	Type	Variable	Type	Variable	Type
Playtime increase player	Cont.	Price	Cont.	Playtime increase player	Cont.
Playtime binary	Bin.	Number of recommendations	Cont.	Number of peers	Cont.
		Number of achievements	Cont.	Euclidean distance	Cont.
		Number of DLCs	Cont.		
		Action	Bin.		
		Adventure	Bin.		
		Indie	Bin.		
		RPG	Bin.		
		Free-to-play	Bin.		
		Strategy	Bin.		
		Simulation	Bin.		
		Multiplayer	Bin.		
		PvP	Bin.		
		Co-op	Bin.		
		Online co-op	Bin.		
		Cross-Platform Multiplayer	Bin.		
		LAN PvP	Bin.		
		Shared/split-screen	Bin.		

Table 12 - Overview of used tables

Genres	*Online capabilities*
Action	Multiplayer
Adventure	PvP
Indie	Co-op
RPG	Online co-op
Free-to-play	Cross-Platform Multiplayer
Strategy	LAN PvP
Simulation	Shared/split-screen

Table 13 - Overview of binary variables

Appendix B

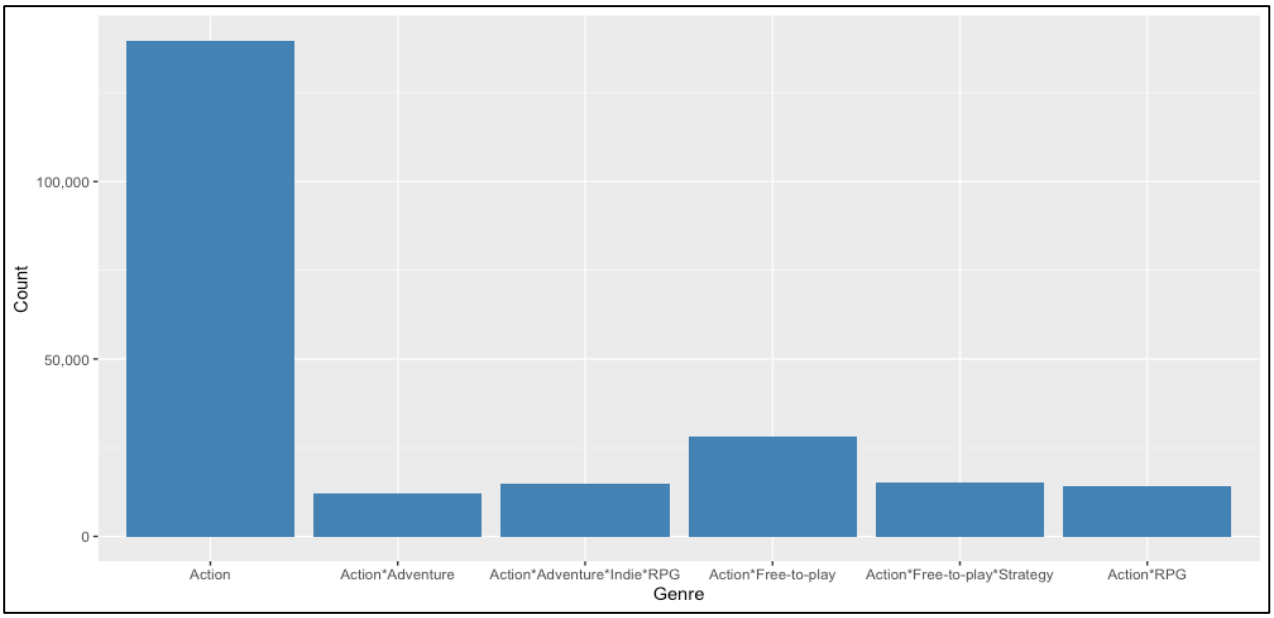

Figure 13 - Most occurring combinations of genres

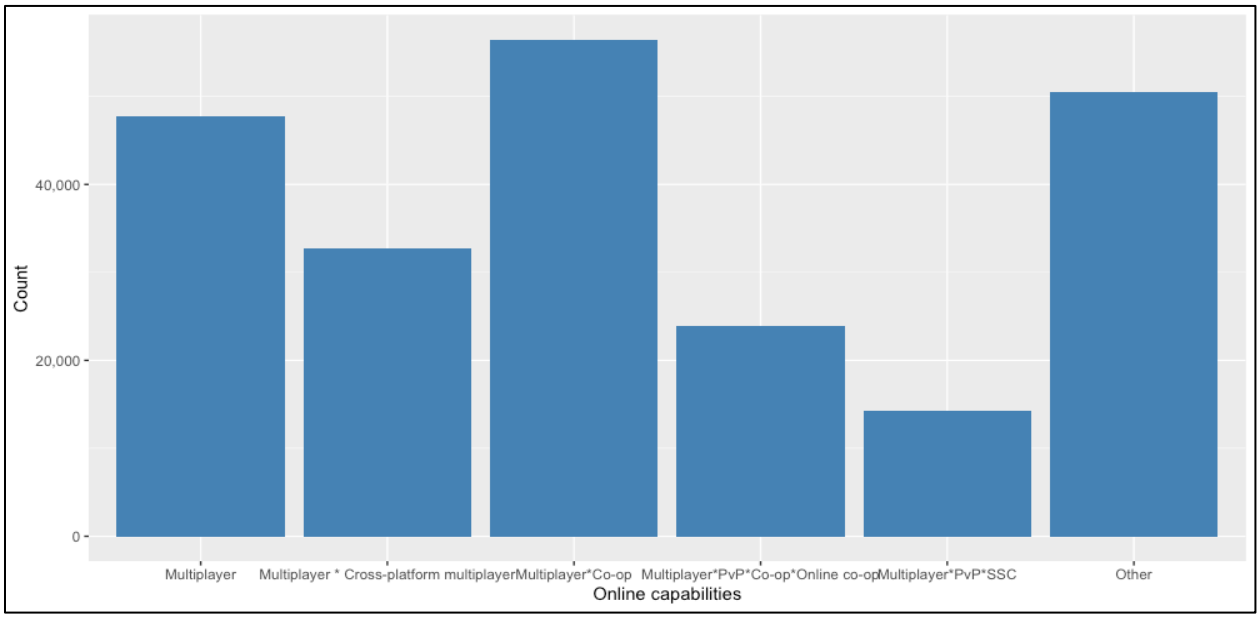

Figure 14 - Most occurring combinations of online capabilities

Appendix C

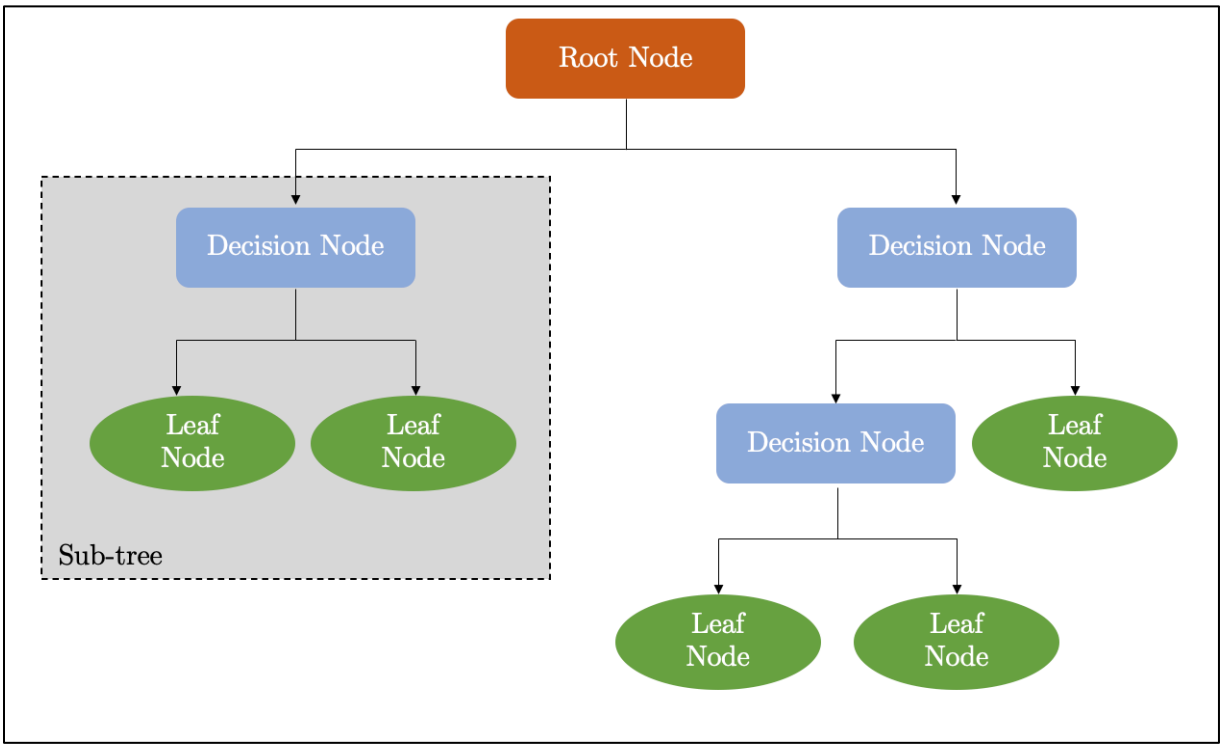

Figure 15 - Visual representation classification tree

Appendix D

	Dependent variable	
Variable	Playtime increase player	
Main effect		
Playtime increase peer	0.611***	(0.018)
Relationship characteristics		
Number of friends	0.216***	(0.006)
Euclidean distance	0.010***	(0.002)
Playtime increase peer * number of friends	0.044***	(0.001)
Playtime increase peer * Euclidean distance	-0.018***	(0.001)
Game characteristics – general variables		
Price	-0.065***	(0.011)
Recommendations	0.395***	(0.006)
Achievements	0.081***	(0.006)
DLC	0.223***	(0.010)
Playtime increase peer * price	-0.078***	(0.003)
Playtime increase peer * recommendations	-0.031***	(0.001)
Playtime increase peer * achievements	-0.010***	(0.002)
Playtime increase peer * DLC	0.017***	(0.002)
Game characteristics – online capabilities		
Multiplayer	0.238***	(0.024)
PvP	0.228***	(0.026)
Co-op	0.240***	(0.024)
Online co-op	-0.354***	(0.043)
Cross Platform Multiplayer	0.346***	(0.034)
LAN PvP	1.166***	(0.128)
Shared/split-screen	0.072***	(0.025)
Playtime increase peer * multiplayer	0.152***	(0.007)
Playtime increase peer * PvP	0.031***	(0.008)
Playtime increase peer * co-op	-0.023***	(0.006)
Playtime increase peer * online co-op	-0.009	(0.010)
Playtime increase peer * Cross Platform Multiplayer	-0.116***	(0.006)
Playtime increase peer * LAN PvP	-0.345***	(0.025)
Playtime increase peer * shared/split-screen	-0.081***	(0.007)
Game characteristics - genres		
Action	-1.744***	(0.046)
Adventure	0.325***	(0.034)
Indie	-0.595***	(0.035)
RPG	0.391***	(0.028)
Free-to-play	2.370***	(0.046)
Strategy	0.558***	(0.033)
Simulation	0.048	(0.062)
Playtime increase peer * action	0.012	(0.009)
Playtime increase peer * adventure	-0.087***	(0.009)
Playtime increase peer * indie	0.093***	(0.009)
Playtime increase peer * RPG	0.020***	(0.006)
Playtime increase peer * free-to-play	-0.193***	(0.010)
Playtime increase peer * strategy	-0.070***	(0.008)
Playtime increase peer * simulation	0.010	(0.013)
Intercept	-2.581***	(0.081)

Table 14 - Outcomes of linear regression model

Appendix E

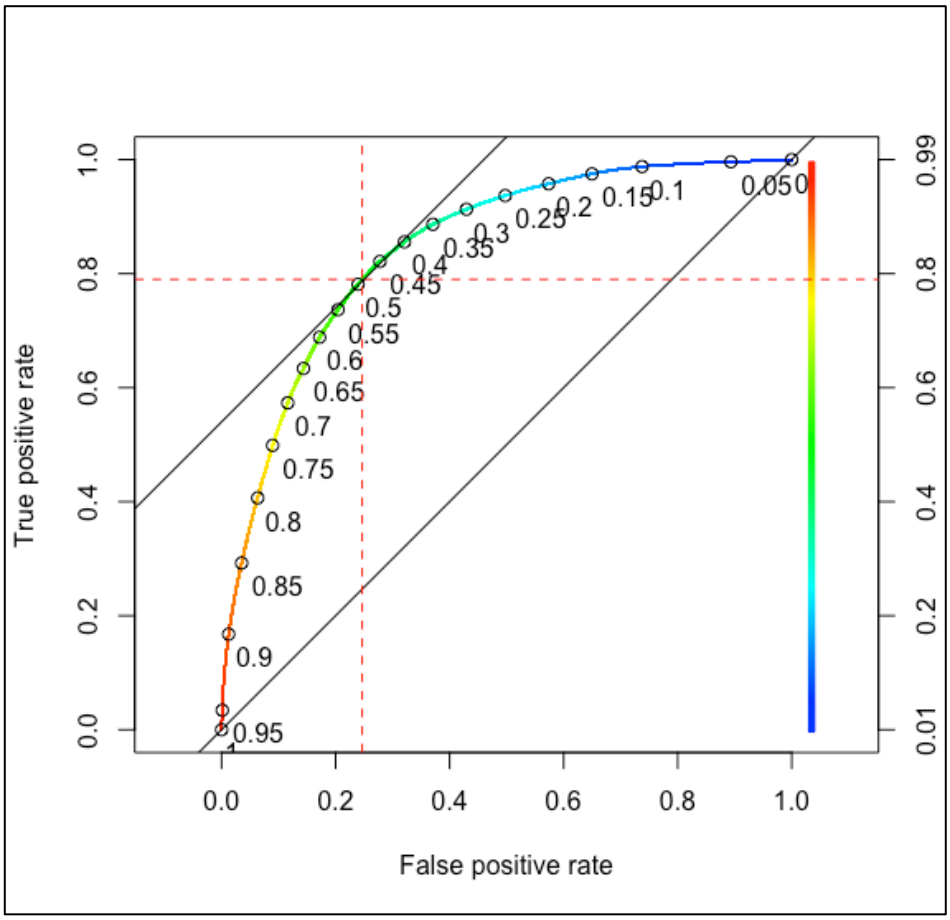

Figure 16 - Visual representation of ROC curve

Statistic	*Value*
Sensitivity	0.79
Specificity	0.75
Threshold	0.49

Table 15 - Overview of performance measures based on the ROC curve