

Eliciting the willingness to pay for mobile virtual goods

Master Thesis Submitted to

Prof. Dr. Ostap Okhrin

Prof. Dr. Brenda Lopez Cabrera

Ladislaus von Bortkiewicz Chair of Statistics

C.A.S.E.- Centre for Applied Statistics and Economics

Humboldt-Universität zu Berlin



by

Polina Marchenko

(521904)

in partial fulfilment of the requirements

for the degree of

Master of Science in Business Administration

Berlin, January 26th, 2012

Declaration of Authorship

I hereby confirm that I have authored this master thesis independently and without use of others than the indicated resources. All passages, which are literally or in general matter taken out of publications or other resources, are marked as such.

Berlin, 26.01.2012

Polina Marchenko

Abstract

This study gives an outline of modern methods of willingness to pay (WTP) elicitation in the realm of private goods. The empirical study applying Contingent Valuation Method (CVM) for WTP elicitation of mobile virtual goods was conducted. Subsequently, the advantages and disadvantages of CVM were discussed. Additionally, the logistic regression analysis and the classification and regression trees (CART) analysis were used in order to distinguish the variables that influence the WTP for mobile virtual goods. Finally, comparison of the predictive ability of both approaches was performed using Receiver Operating Characteristic (ROC) Analysis.

Keywords: Willingness to pay, Contingent Valuation Method, CART, ROC, logistic regression

Contents

1	Introduction	1
2	Previous research	4
3	Digital virtual goods	7
3.1	Mobile virtual goods	10
3.2	Virtual goods and music industry	11
4	Design of empirical study	13
4.1	Product description	14
4.2	Sample description	16
5	Willingness to pay prediction with logistic regression	22
5.1	Logistic regression model	22
5.2	Fitting the logistic regression model	23
5.3	Interpretation of the logistic regression parameters	24
5.4	Model selection	25
5.5	Empirical results	26
6	Willingness to pay prediction with CART	32
6.1	Growing the classification tree	32
6.2	Tree pruning methods	34
6.3	Empirical results	36
7	Model performance assessment metrics	44
7.1	Confusion matrix	44
7.2	Receiver Operating Characteristic Analysis	46
7.3	Empirical results	47
8	Conclusion	53
	Appendix	57

List of Figures

- 3.1 Revenue shift from in-app advertisement to in-app purchases 11
- 3.2 Music applications in the Apple App Store 12

- 4.1 Fourfold plots of association between gender, age and smartphone 17
- 4.2 Conditional plot virtual goods, gender and age 18
- 4.3 Music exploration preferences 19
- 4.4 Willingness to listen and share music with friends 20
- 4.5 Willingness to listen and share music with professionals 20
- 4.6 Willingness to listen and share music with unknown people 20
- 4.7 Hypothetical WTP versus calibrated WTP response rates 21

- 6.1 Cost-complexity pruning tree sequence 38
- 6.2 Pruned classification tree for the unlimited following slot (ww1) 40
- 6.3 Pruned classification tree for the advanced profile (ww2) 41
- 6.4 Pruned classification tree for the extended range (ww3) 42
- 6.5 Pruned classification tree for the exclusive live music streams (ww4) 43

- 7.1 Important regions and points of ROC graphs 47
- 7.2 ROC for the unlimited following slot (ww1) 49
- 7.3 ROC for the advanced profile (ww2) 49
- 7.4 ROC for the extended range (ww3) 50
- 7.5 ROC for the exclusive live music streams (ww4) 50

List of Tables

2.1	Calibrating results	5
3.1	Mobile music applications	12
4.1	Personal characteristics	14
4.2	Dependent variables description	15
4.3	Independent variables description	16
5.1	Design variables	24
5.2	Cross-classification table	25
5.3	Assignment of design variables	27
5.4	Best logistic model for the unlimited following slot (ww1)	28
5.5	Best logistic model for the advanced profile (ww2)	29
5.6	Best logistic model for the extended range (ww3)	29
5.7	Best logistic model for the exclusive virtual ticket (ww4)	30
5.8	Likelihood-ratio test results	31
6.1	Significant variables of maximum and pruned tree	39
7.1	Confusion matrix	44
7.2	Confusion matrix with and without class skewness	45
7.3	Significant variables used in CART and logistic regression	48
7.4	Most significant differences between learning and test samples	51
7.5	Willingness to pay rates in learning and test samples	51

1 Introduction

'You can have anything in this world you want, if you want it badly enough and you're willing to pay the price' Mary Kay Ash (Founder of Mary Kay Cosmetics)

Price is one of the most critical characteristics, often decisive for successful transactions of goods or services. Willingness to pay (WTP) value stands for the maximum reservation price a person is willing to pay in order to receive goods or services or in order to avoid some undesired phenomenon. Information about WTP is required, in order to determine the optimal pricing policy for goods or services. In case of private goods, it is important for the survival and success of the company, whereas the appropriate price policy for public goods has an influence on the different aspects of national welfare. Revealing such information is neither straightforward for scholars or managers. It is hard to argue that everything in the world has its price, our question is: *"How price can be determined?"*

There are numerous approaches to measure the WTP. According to the framework given in Breidert et al. (2006), on the highest level, the methods can be divided according to the source of data used. There are three main sources of the WTP values: methods based on *actual sales data*, e.g. from customer panels, *simulated preference data*, revealed in experiments or auctions and *stated preference data*, represented by direct and indirect surveys.

Further, the approaches can be distinguished between the open-ended elicitation methods, such as Becker-DeGroot-Marschak (BDM) auction, Becker et al. (1964) and Vickrey (second price) auction, Vickrey (1961) and, the closed-ended elicitation methods, such as the contingent valuation method (CVM). The choice of the method depends on the purpose of the researcher. The difference between these methods will be clear after looking at the research questions. Open-ended elicitation methods deliver the answer to the question: *"What is someone's maximum willingness to pay for the offered good?"*, whereas the individuals in the closed-ended elicitation methods are confronted with the question: *"Would someone be willing to pay the stated price for the offered good?"*.

Hence, in the open-ended auction a respondent reveals his reservation price for a good. In contrary, the CVM employs the dichotomous choice mechanism, which means that a participant rather compares his reservation price for a good with the offered price. As a result of the open-ended auction, the range and the average price for the good can be

estimated. Since the answer format in the closed-ended method is reduced to a simple *yes - no* decision, one can only test whether the specified price level is appropriate for the good or not.

Moreover, product type as well as budget and time constraints determine the method. The main difficulty of price research in the private goods segment lies in the difference between large and small companies. Large companies possess budgets for market research purposes. They can afford to spend considerable amounts of money in order to conduct large scale studies for their new product. The reality looks different for small start-up firms, which are those that drive innovation. Such firms have flat hierarchies, but no marketing budgets. Hence, the pricing policy is often managed by using a "*trial and error*" process.

Such a situation is also typical for small innovative firms in the digital goods sector. Digital virtual goods, in the past categorised as "*money for nothing*", nowadays become a part of the daily routine of the *Facebook* generation. There is no doubt that price setting is also the key activity of businesses in this field. To our knowledge, there is a lack of comprehensive studies on the subject of WTP patterns for digital virtual goods, and the main purpose of this work is to fill this information gap

Open-ended auctions conducted are typically characterised by the physical presence of people and the auction subject being a material private item. On the one hand, the advantage of such auctions is that the real money-good transfer takes place, on the other hand, the drawback is that the auction situation is not typical for a common consumer and might be misinterpreted, leading to biased results, see Skiera and Revenstorff (1999). Furthermore, the auction procedure is difficult to realise and yields high costs as well as the presence of an item to be sold.

Taking into account the possible costs of auction as well as the specifics of the virtual goods, the closed-ended elicitation method CVM is considered as the most appropriate for determining WTP and these considerations were used in this study. A questionnaire based method is common for market research practices and is widely applied as a cost-saving method by different companies. Nevertheless, one of the major limitations of this method is the hypothetical nature of the elicited WTP values. Hence, by additionally applying the *ex post* calibrating procedure, we aim to mitigate the hypothetical bias, which is characterised as the difference between hypothetical and real WTP. In order to provide reasonable conclusions, we compare values of hypothetical and calibrated WTP with the market benchmark.

The subject of this study is a mobile application, which represents a disruptively new way of music consumption, i.e. exploration. The service is available for free for the user, whereas the additional options (virtual goods) are offered at extra charge. By applying the CVM we aim to elicit the WTP for virtual goods in the music sector, in order to

give a notion of whether this method can be offered as a reliable cost-saving method of WTP elicitation for small firms in the mobile industry or rather not.

Furthermore, applying the logistic regression to our data, we aim to distinguish the factors which influence the WTP for digital virtual goods in the music sector. A popular method for tree-based regression and classification called CART was used as a non-parametric alternative to logistic regression. The Support Vector Machines classification approach was also examined, but further research has revealed its irrelevance, when independent variables are of discrete type, i.e. dichotomous and categorical as is the case in our survey. Additionally we provide the model performance assessment analysis by the means of Receiver Operating Characteristic (ROC) analysis in order to find out whether two chosen classification concepts are able to deliver reasonable predictive ability of willingness to pay patterns on hypothetical data in the form of in-sample and out-of-sample predictions.

The outline of this work is as follows: Chapter 2 contains an overview of the previous studies about the willingness to pay using the contingent valuation method and the hypothetical bias mitigation calibrating methods. Chapter 3 provides a description of both digital and virtual goods, as well as the peculiarities of mobile virtual goods and their importance for the music industry. The following Chapter 4 contains information about empirical survey design, product description and provides descriptive statistics of the sample. The theoretical interpretation of the logistic regression model and empirical results of the willingness to pay prediction with logistic regression are given in Chapter 5. Chapter 6 offers the theoretical background for the CART model and empirical results of the model application. Chapter 7 is devoted to the model performance assessment analysis, describing a ROC comparison evaluation of both models. Finally, concluding remarks are given. The Appendix contains a script of the online survey at *limesurvey.com*.

2 Previous research

The contingent valuation method (CVM) was originally developed by Robert Mitchell and Richard Carson in 1989 with the purpose of measuring the willingness to pay for environmental changes, Mitchell and Carson (1989). At the beginning the CVM was used to determine the price level for the non-marketed goods. Later on the CVM was also used in studies with private goods, Johannesson et al. (1998).

One of the major limitations of this method is the hypothetical nature of the revealed willingness to pay. CVM surveys are hypothetical in both payment and provision of the good. Therefore, many economists argue whether individuals' responses in a hypothetical setting reflect their actions in the real decision situations and whether these hypothetical values can be used as a notion for price setting in market practice. In spite of these disadvantages, CVM questionnaires continue to play an important role in market research.

Hypothetical decision making is assumed when there are no consequences associated with individual's response. On the contrary, the real purchase decision obliges individual to pay the stated price. The most prominent works in this field are: Johannesson et al. (1998), Blumenschein et al. (1998), Harrison and Rutström (2008), Johannesson et al. (1999) and Blumenschein et al. (2008).

The discrepancy between hypothetical WTP and actual purchase decisions has a name: *hypothetical bias*.

Hypothetical bias occurs if values found in a hypothetical context significantly differ from the results elicited in a real market situation. Experiments carried out by Cummings (1997) and replicated by Johannesson et al. (1998), confirmed the overestimation of real purchase decisions by the hypothetical answers given in CVM. According to these findings, one assumes individuals to be biased by the hypothetical nature of the experiment, since they know that, independent of their decision, they would not have to spend money. The discussion triggered by these results has started the new research wave, centred on the possibility of mitigation of the hypothetical bias and producing unbiased WTP estimates also using the CVM study.

It is important to distinguish between *ex ante* and *ex post* calibration methods. The *ex ante* method is, for example, *cheap talk*, the purpose of which is to make respondents aware of the hypothetical bias before making a purchase statement, in order to encourage the decision making as if there were real economic consequences. This approach was

successfully applied by Cummings and Taylor (1999) in the CVM study with environmental goods. However, the robustness of this calibrating approach was not supported by a later study using an auction design with private goods (sports-cards), List and Lucking-Reiley (2000). Another study with private goods (art prints) by Loomis et al. (1996) although, suggested cheap talk to reduce the hypothetical bias, the results were not statistically significant.

Ex post methods aim to calibrate the responses after the WTP statements are done. There are two known calibration methods of this type: one implementing the 2-levels certainty scale and another offering the 10-levels certainty scale. In both cases, after answering the WTP question, individuals are confronted with the follow-up certainty question: "How sure you are about buying the good X at the price Y?".

According to the first method, two possible answers exist, "definitely sure" and "probably sure", whereas in a 1-10 scale one can decide from "very uncertain" to "very certain". By using this procedure, a researcher is able to classify the hypothetical buyers into two categories. People who answered "yes" in a hypothetical WTP question and are "definitely sure", can be considered as *buyers* in the real situation. Whereas respondents who answered "yes" to the WTP question but are "probably sure" about that, can be identified as *non-buyers*. Individuals who gave negative answer in hypothetical situation are considered as *non-buyers* independently on their certainty level. Table 2.1 represents the interpretation of the calibrating procedure.

		Certainty level	
		definitely sure	probably sure
WTP	yes	real buyer	non-buyer
	no	non-buyer	non-buyer

Table 2.1: Calibrating results

There is different, somewhat contradicting evidence of how successful these methods are in practice. In his survey Johannesson et al. (1998) applied the 2-levels certainty scale. He tested the hypothesis suggesting "definitely sure yes" responses correspond to the "real yes" responses. The hypothetical bias is tested by the calculation of the discrepancy between the proportions of hypothetical and real yes responses. The study confirmed the hypothetical bias, but revealed that this calibrating method tends to significantly underestimate the "real yes" responses. Though, the null hypothesis was rejected.

In contrast, the later studies of Blumenschein et al. (1998) and Blumenschein et al. (2008) could not reject the null hypothesis of no difference between the proportions of "definitely sure yes" responses and "real yes" responses. These studies were carried out

with private goods.

In another experimental study about WTP for public goods Champ et al. (1997) used the 1 – 10 certainty scale. The existence of the hypothetical bias was also stated. In his work Champ considered only "very certain" answers to correspond to the real purchase decisions, but did not find any significant evidence that could predict the "real yes" responses.

Johannesson et al. (1999) used data from within sample comparisons of the two previous experiments of Blumenschein et al. (1998) and Johannesson et al. (1998) and was able to estimate the statistical bias function. He applied the 1–10 certainty scale as a calibrating method. Moreover, his findings revealed that "real yes" responses can be accurately estimated by the calibrated hypothetical responses. Herewith the null hypothesis of no significant difference between hypothetical "definitely sure yes" responses and "real yes" responses could not be rejected.

Blomquist et al. (2009) study included data sets for three different health programmes, comparing the effectiveness of 2-levels and 10-levels of certainty scales in mitigating the hypothetical bias. The experiment confirms that "definitely sure yes" corresponds to the "yes responses of the 8th certainty level" and both calibrating techniques can be an indicator for "real yes" responses. Generally, the results of the studies suggest calibration to be appropriate to filter out individuals who will really pay from those who only say they will.

In this work the *ex-post* method was preferred to *ex-ante* method, because the survey was performed online and without personal contact with the respondent. For these reasons the *cheap-talk* method was classified as lacking convincing power as well as being time-consuming, therefore inappropriate for this survey design.

While employing the calibrating methods described, the nature of the good might be an important factor to consider. As suggested in the meta-analysis of List and Gallet (2001), hypothetical bias is considerably higher for public goods. The intuition behind this conclusion is that people are usually more familiar with the context of private goods and therefore are able to provide evaluations containing less errors. However, the results of the further extended meta-analysis by Little and Berrens (2004) did not support previous findings and rather suggest the nature of the good does not have an influence on the disparity between hypothetical and real values.

To our knowledge of the previous research in the field of contingent valuation analysis, digital virtual goods have not been a subject of an investigation yet. The explosive proliferation of virtual goods in the few last years creates both opportunities and challenges for companies. Considering the lack of attention to this field, the study investigating the purchase patterns for this kind of goods is of great interest.

3 Digital virtual goods

The subject of current study is a mobile digital service, i.e. mobile application, which by its nature is a disruptive innovation, since it represents an absolutely new way of music consuming. The service is available for free for users, whereas the additional options (virtual goods) are offered at an extra charge. Our purpose is to apply the CVM for assessing the WTP for virtual goods, in order to give a notion of whether this method can be offered as a cost-saving method of WTP elicitation for small firms in the mobile sector.

Disruptive innovations in the private goods market can be, to some extent, compared to such non-marketed goods as health, safety and environment, because for all of them, markets do not exist. *Disruptive innovations* are also called *discontinuous innovations*, because they push the progress into the unexpected earlier directions. A good example of disruptive innovation is the business model of the American low-cost air carrier *Southwest Airlines*, since they drastically changed flight ticket price concepts. *Southwest Airlines* managed to cut their prices by the introduction of an additional charge for luggage and meals on the board.

No doubt, some of the most influential disruptive innovations in the digital world in the last century were Voice over IP (VoIP), standardised by *Skype* for the global market; touch screen technology, originated by *IBM* and effectively merchandised by *Apple* and, last but not least, *iTunes* music online store.

The literature review about digital and virtual goods provides a mixed explanation for these types of goods, because it is not simple to distinguish between these goods. For this reason, in the following passage we try to summarise the existing definitions in order to provide our understanding of digital virtual goods.

According to Stelzer (2004), digital goods are non-material goods, which can be developed, sold and used by and within the information systems. Digital goods can be categorised by the degree of digitalisation. Therefore, there are three types of digital goods: completely digital goods, for instance, software downloaded from the internet or music stream; digital goods on tangible mediums, such as software delivered with manual and digital goods with consultation, for example, software which is sold in packages within a seminar by professional consultants.

Mandy Salomon (Swinburne University of Technology, Australia) provides a very good definition of the virtual goods, although the researcher uses "digital goods" heading: "A

3 Digital virtual goods

digital good is really just a piece of code, which has been turned into something that's graphically seen as being a good of some sort. It doesn't have any intrinsic value but it has a perceived value by the user. In other words, you can be looking at a bunch of roses, or you can be looking at a hat, or some sort of attractive garment that might be good for your online persona, your avatar. But equally a virtual good can be a service; it can be something that makes you do something better in a virtual game. [...]"

Digital virtual consumption differs from material goods consumption since the object of consumption does not have material substance and cannot be used in material reality. As suggested by Denegri-Knott and Molesworth (2010) digital virtual to be categorised as "*liminal*" - hybridisation between the imaginary and the material world. Material dimension includes PCs, smartphone screens, headphones and always embodies an end user.

The imaginary element of the digital virtual goods consumption (DVC) according to Denegri-Knott and Molesworth (2010) is based on four main functions:

- stimulates the consumer desire in the virtual space, which also has a stimulative effect on material consumption;
- enacts consumers daydreams, ownership of the different products in real life may not be possible due to budget constrain, whereas in the virtual space, for far less money consumers live their daydreams of wealth and status;
- turns consumer fantasies into reality (although virtual reality), it is possible to become a super hero, who does not exist in the real world, *however, one is not a super hero, but one acquires a feeling that one is.*;
- stimulates experimentation, meaning that one can adopt different social roles without any negative consequences.

Also, according to the historical timeline, digital goods evolved in a form of different software programs, with the emergence of personal computers in the latter part of the 20th century; whereas the first virtual goods were introduced only in the late 80's.

That is why taking into consideration both these facts, we suggest digital good as being a generic term, which contains the definition of a virtual good within it. According to the classification by Stelzer (2004), a virtual good can be defined as a completely digital good, since it exists only in digital form.

The history of the virtual goods consumption begins in 1985, in the year when virtual goods were first introduced by the virtual 2D environment, *Habitat*. At that time the virtual goods used to be bought for virtual currency, which itself was free distributed among players. Already in 1999 the revolution in virtual trading took place, as the

virtual items from other popular games *Ultima Online* and *EverQuest* were traded for the real money at *eBay* auctions.

Nowadays the idea of operating with virtual goods has spread beyond its origins in massively multiplayer online role-playing games (MMORPG) and found its future development in online social communities. The most popular social network, *Facebook*, with over 800 Million users worldwide in January 2012, facebook.com, and 22 Million users in Germany in January 2012, allfacebook.de, benefits from selling virtual goods. Escalating revenues of *Zynga*, the largest producer of social games on *Facebook*, are evidence of boom in social gaming. Inside Network, a research and media organisation, predicts its revenues to reach 500 million US Dollars in 2011. *Zynga's* games are free and its revenues come mainly from selling virtual goods that players can obtain within games. Although *Zynga* is an absolute leader in the social games industry, according to the company's own statement less than 5% of their players are actually paying players, Reuters (2011). Whereas Wedbush Securities analyst Michael Pachter suggests the industry average monetization level to be under 2%. Paul Verna, analyst of *eMarketer*, is more optimistic about the U.S. social games market, he estimates that paying gamers make up 6% of all social game players in the U.S.

According to the value framework introduced by Sheth et al. (1991) there are three pertinent dimensions of customer consumption values: functional value, emotional value and social value. All three dimensions were proved to be key influencers on consumers behaviour.

Functional value incorporates such attributes as reliability, durability and price. Emotional value stands for the product's capacity to arouse feelings. Social value of the product is made of such attributes as symbolic meanings, social relationships and own identity. There is no doubt that all of these characteristics are not less pronounced in the consumption of virtual goods.

Another explanation for the individual's consumption is given by Jeremy Liew (Light-speed Venture Partners), who suggests people buy virtual goods for the same reasons that they buy goods in the real world: first, to be able to do more, for instance, new personal computer versus new levels in the game; second, to establish and maintain social contacts, for instance, gifts in real life versus gifts on *Facebook* and, third, to express their personality, for instance, new clothes in real life versus avatars items in the game. In alignment with previous considerations, and also according to Schneider (2008) - the world's first combined e-commerce and advertising platform for virtual goods, three types of virtual goods can be defined:

- Vanity items - items that allow players to customise an avatar.
- Functional items - items used to progress in a game.

- Social items - items to be gifted to other users.

Economists suggest that what was previously considered to be fiction can actually be analysed as goods in an economic sense, Castronova (2002), Lehdonvirta et al. (2009). Meanwhile, when goods are labelled "virtual", it is not meant any more that these goods are less *real*, they are rather computer-mediated, Lehdonvirta (2008). Although virtual objects are technically speaking not more than a series of pixels, they deliver far more intrinsic value for the user. Nowadays it is out of question that people spend money for the virtual goods as well as they do for material goods, that is why it is worthwhile to pay attention to this market.

3.1 Mobile virtual goods

Behind virtual communities stand high profits, and new ways of games monetisation evolve for the purpose of profit maximisation. In order to better integrate the purchasing decision into the game environment and so to increase the number of purchases, *RubyCoins* has developed the inGame payment or micro-transaction mechanism, which enables the exchange of real money for virtual goods within the online game.

The trend of total mobilisation of the world society, increasing amount of smartphone users worldwide shift the virtual goods consumption into the mobile space and empowers mobile applications. In the meantime, a user can be engaged in social interaction whenever and wherever they wish, non-stop and on-the-go. This flexibility facilitates mobile virtual goods consumption and brings it to the next level.

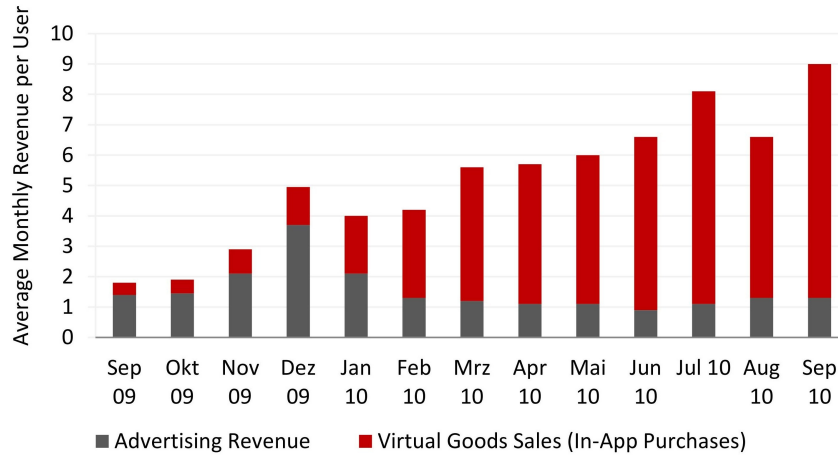
Micro-transactions mechanism, introduced in 2009 for Apple iOS, allows users to buy goods in application (in-app). This technology also enables mobile services other than games to increase their revenues due to in-app virtual goods distribution.

The majority of applications are based on the Freemium model, which implies the core product to be free and a premium content to be paid. For this reason, it was assumed that advertising would become the largest part of the revenue streams, however, the survey by analyst firm *Flurry* reveals the leading role of virtual goods for the application monetisation, see Figure 3.1.

Average revenue per user (ARPU) for virtual goods surpassed advertising ARPU and, moreover, has an upward trend.

The results of the Magid Media Futures 2010 Wireless and Consumers Report, Magid (2010) convey the importance of the mobile market for virtual goods. According to the report, around 23% of the American population own smartphones, which is about 122 million people, and 45% of smartphone owners are engaged in mobile gaming, which adds up to about 55 millions. A total of 168 million Dollars were spent on mobile virtual goods in 2009 by Americans.

3 Digital virtual goods



Source: Flurry

Figure 3.1: Revenue shift from in-app advertisement to in-app purchases

Today the idea of selling virtual commodities has spread beyond the gaming industry and is about to become a successful monetization model in other areas, particularly those where social interaction is a key element. Music is a tool for social exchange, that is why the implementation of the virtual goods idea within the mobile music application is considered as a lucrative business model.

3.2 Virtual goods and music industry

The main purpose of the new and existing music services is to provide consumers with legitimate alternatives to piracy. Nielsen (2010) suggests about one quarter of active internet users in Europe visit illegal unlicensed file sharing sites monthly, which causes great losses for the music industry.

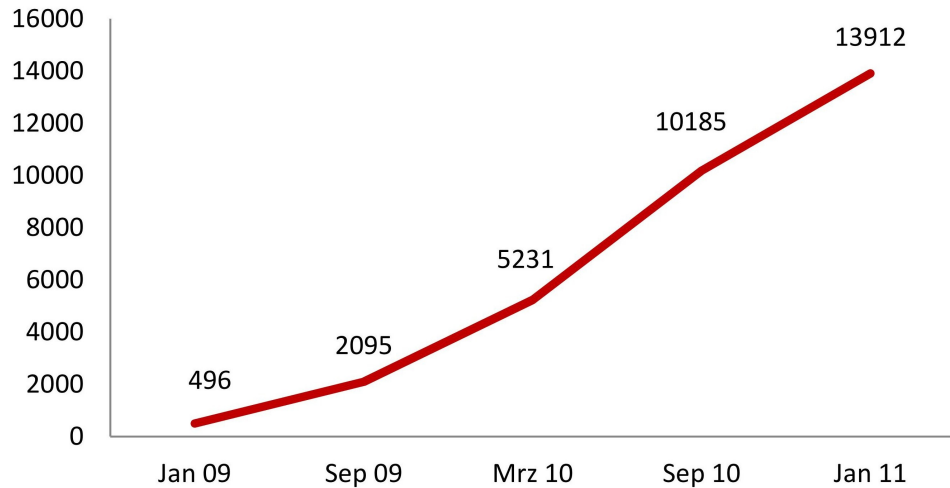
According to IFPI (2011) *because it is free* argument appears to be the major motive for illegal music downloading as opposed to other factors such as better choice, convenience or quality of service of the legal providers.

Mobile music applications broadly implement the *Freemium* business model, with two kinds of offering to consumers, free and premium. Such services represent the upcoming channel for legal music exploration due to its broad music offers and relatively low costs. In the Report of Nielsen (2010) music applications are classified into four categories, mainly artists' applications, music discovery applications, streaming applications and live concert applications. There is evidence that consumers from Europe prefer music discovery applications the most, though 45% of these name this type of app as the most interesting for them. This tendency is also relevant for consumers in other world regions. About 35% of Europeans name both artists' applications and streaming applications as

3 Digital virtual goods

the most relevant to their interests. Live concert applications are indicated as interesting by more than 25% of European customers.

Mobile music applications represent service, which replace and improve older methods of music distribution. The rise of the amount of music applications in the Apple App Store, according to Informa Telecoms & Media agency, is clear proof of the previous statement, see Figure 3.2.



Source: Informa Telecoms & Media

Figure 3.2: Growth of the amount of music applications in the Apple App Store

Although it should be clear that the most of these applications often duplicate the functionality of others and the majority of them cannot compete. The most important and prominent music applications in Germany are listed in the Table 3.1.

Name	Value proposition	Monetization
Simfy	music streaming	advertising or usage fee
Last.fm	recommendation radio	usage fee
Soundcloud	record, stream and store audio	freemium
TuneWiki	music streaming with lyrics	freemium
Shazam	identifying music tunes	freemium

Table 3.1: Mobile music applications

The music application, described in this work, might have a chance to succeed, because it opens a new "social music" market and does not compete in an established one.

4 Design of empirical study

Following the theoretical considerations discussed above, we perform the willingness to pay survey for mobile virtual goods.

The survey questionnaire contains a total of 22 questions. This questionnaire was distributed online via a student forum at the Technical University Berlin and via a student mailing list at the Humboldt-Universität Berlin. Altogether 625 usable completed answer sheets were collected, all of the questions were set to be obligatory for respondents, so that our data sample has no missing values.

Each respondent received a hypothetical dichotomous questions followed by a 2-level certainty question concerning previously stated WTP of the feature Y , $Y = (Y_1, Y_2, Y_3, Y_4)^T$ at a price of € 0.79 for each of the first three features and € 2.29 for the last one. In contrast to Blumenschein et al. (2008) study, the certainty question was received by all subjects, and not only by the subjects, who answered yes to the willingness to pay question.

Altogether four questions about WTP for different features were asked, this practice had never been used in previous studies. Normally the subject was confronted only with one WTP decision. Hence, testing 4 features should give us a notion about the level of interest for different features. The last feature is offered at the reasonably higher price, because it delivers the most visible functionality and is labelled as "exclusive", which should be transferred into the price level. In such a way the aim was to gain knowledge in valuation of different features of the product.

Previous to WTP questions a short description of the service in general as well as a description of the features, were given. Since the script was time consuming for individuals, the questions were kept as short as possible. For the same reason, we did not use any *ex ante* calibrating methods. The survey sheet can be found in Appendix. Furthermore, in order to distinguish the factors which influence the willingness to pay for virtual goods, the individuals received questions about their personal characteristics, summarised in Table 4.1.

We then ran the logistic regression and conducted the non parametric CART analysis of the collected data. The willingness to pay questions were considered as dependent variables, whereas all other personal data variables were treated as independent variables. Consequently we used ROC analysis in order to visually depict the performance and performance trade-off of both classification models.

Category	Characteristic
Demographical data	gender age
Smartphone usage patterns	cellphone type monthly budget for mobile applications
Social music affinity	monthly budget for digital music music exploration type willingness to listen music willingness to share music
Social games patterns	social games experience engagement level purchase experience of virtual goods

Table 4.1: Personal characteristics

As the third part of the survey the real WTP decisions should have been tested with a group of individuals who had been already using the application for a short period of time. Unfortunately, at the early stages of this work, it was revealed that this aspect was not possible to complete due to technical immaturity of the mobile application.

Hence, our purpose is to investigate to which extent the stated hypothetical WTP is able to predict the decisions under real market circumstances for the complex case of innovative digital (virtual) products. Moreover, within this study we wanted to compare: hypothetical WTP and WTP values, adjusted by the certainty question, against a market benchmark.

The experiments cited above show that there is evidence that hypothetical "definitely sure yes" responses mitigate hypothetical bias compared to hypothetical "yes" responses without certainty statement calibrations. Despite this fact, there is no theory supporting this experimental evidence. For this reason, the results of previous studies on private and public goods cannot be generalised for the use for virtual goods, offered within mobile smartphone applications.

4.1 Product description

wahwah.fm is a location based music application for iPhone. The core functionality of the application includes:

- possibility to listen in real time what other users are listening to;
- possibility to create private radio station and make it available for public use.

Any user of the application, from the music community who finds the music one streams to their liking can become a listener. Due to the technical know-how of the provider, this service is not a file-sharing platform, but a legal music service for music exploration. The core functions are free, the additional features are offered as in-app items:

1. *Unlimited following slot* (functional item) can be explained as the ability to get access to the favourite broadcaster's music streams, independent of his or her location. This feature might be compared to gaining more functionality in the virtual games and it also reflects consumer's desire of having many friends.
2. *Advanced profile* (vanity item) offers a possibility to customise one's own profile and make it more prominent than others. This feature reflects the function of avatar, and may be interpreted as a demonstration of status or belonging. It turns a wish of being a famous Dj into reality that is - true for the virtual community.
3. *Extended range* (functional item) feature is similar to the first feature and gives a possibility to explore unknown broadcasters in chosen places in the world. This can be understood as an analogy to the new level in virtual games, as one can open the secret area and get an access to other personal music stations not previously available. Moreover, the imaginary presence in other cities may reflect a consumer's daydream of travelling and caters to the interest to other cultures.
4. *Exclusive virtual ticket* (functional item) enables attendance of a real music event digitally. This feature can be interpreted as a demonstration of status, since the exclusivity is underlined in the description and in the premium price.

Four in-app features described above correspond to the four dependent variables, these are treated separately in the further analysis and are listed in Table 4.2 below.

Dependent variable	Abbreviation	Category
Willigness to pay for:		
Unlimited following slot	ww1	no purchase / purchase
Advanced profile	ww2	no purchase / purchase
Extended range	ww3	no purchase / purchase
Exclusive virtual ticket	ww4	no purchase / purchase

Table 4.2: Dependent variables description

4.2 Sample description

The list of independent variables with their abbreviations as used in R is given in Table 4.3.

Independent variable	Abbreviation	Category
Gender	gen	male / female
Age	age	18 - 24 / 25 - 31 / 32 - 45
Mobile operation system	os	no smartphone / smartphone
Budget for mobile applications, monthly	bapp	0 € / <5 € / 5 - 10 € / >10 €
Budget for digital music, monthly	bmus	0 € / <5 € / 5 - 10 € / >10 €
Explore new music via:		
Internet	int	no / yes
Radio	rad	no / yes
TV	tv	no / yes
Friends	fr	no / yes
Willingness to listen the music of:		
Friends	lfr	no / uncertain / yes
Acquaintances	lac	no / uncertain / yes
Social network contacts	lsc	no / uncertain / yes
Professionals	lpr	no / uncertain / yes
Unknown people	lun	no / uncertain / yes
Willingness to share the music with:		
Friends	sfr	no / uncertain / yes
Acquaintances	sac	no / uncertain / yes
Social network contacts	ssc	no / uncertain / yes
Professionals	spr	no / uncertain / yes
Unknown people	sun	no / uncertain / yes
Social games		
Experience with social games	soga	no / uncertain / yes
Engagement level with social games	enlev	no / low / middle / high
Purchase experience with virtual goods	vigo	no / yes

Table 4.3: Independent variables description

According to gender distribution our sample population is quite heterogeneous and consists of 61.4% female and 38.6% male respondents.

The age structure of the sample is composed of four age groups: 18 – 24 years, 25 – 31 years, 32 – 38 years and 39 – 45 years. It should be mentioned that the distribution is skewed in the direction of the younger respondents, so the majority of the sample, approximately 88.2% are representatives of the two younger groups, whereas only seven individuals represent the oldest group. Since the group of 39 – 45 years old respondents is too small, for the further descriptive analysis it was merged with the group of 32 – 38 years old.

4 Design of empirical study

32% individuals in the sample are smartphone users, which is a considerably higher rate than Germany's average of 23%, Block (2011). Both age distribution and smartphone usage rate can be explained by the fact that the questionnaire was distributed primarily between students, who are more tech-savvy than other social groups.

By computing the odds ratios for the dependence between age, gender and smartphone usage, we can conclude that only in the age group "25–31" years old, there is a significant association between gender and preference for smartphones. Though the likelihood to possess a smartphone rather than a standard cell phone rises from women to men, since the odds ratio is 1.91 with a 5% significance level. This is illustrated on the fourfold plots, see Figure 4.1. On the fourfold plots the area of the quarter circles is proportional to cell frequency and the rings of adjacent quadrants represent the odds ratios, which overlap only if the observed counts are consistent with the null hypothesis of non association between variables, gender and smartphone.

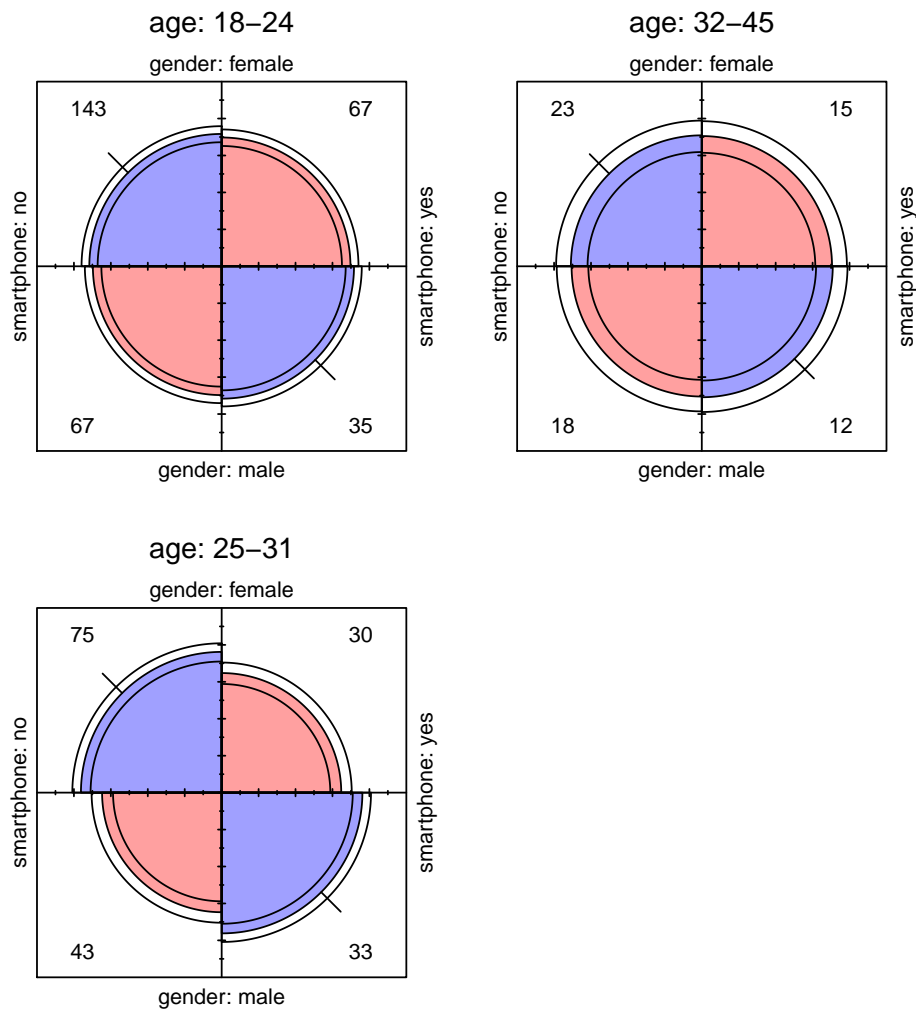


Figure 4.1: Fourfold plots of association between gender, age and smartphone

4 Design of empirical study

The Figure 4.2 visualises the probability of acquiring virtual goods, given gender and age. For three groups we can conclude that the probability of purchasing virtual goods is higher for the male respondents.

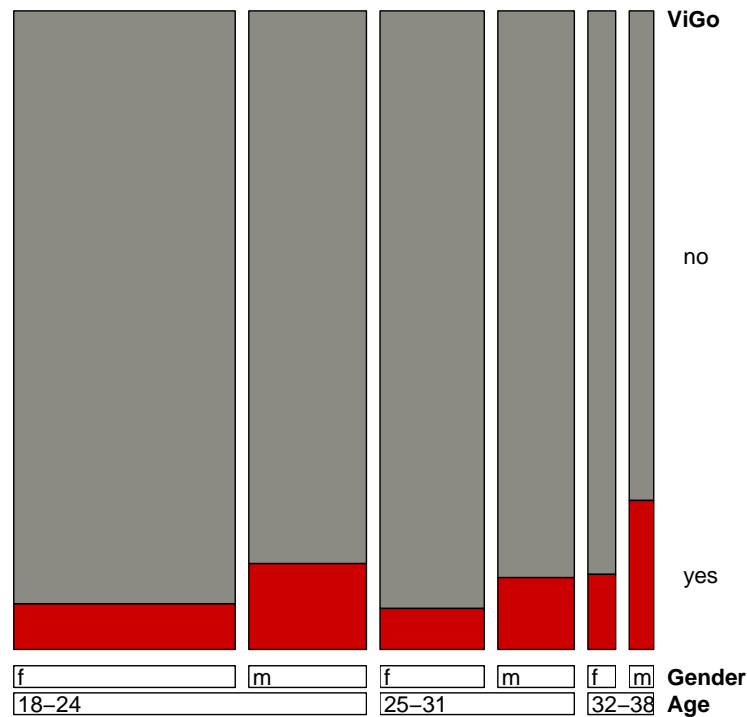


Figure 4.2: Conditional plot of the probability of purchasing virtual goods, with gender and age as conditional variables

Nearly 60% of all respondents spend no money on mobile applications, the same tendency is also true concerning expenditure for the digital music.

The majority of the respondents, who spend money for applications and/or digital music invest less than € 5 per month. Nevertheless around 12% of respondents spend between € 5 and € 10 monthly for applications and/or music, whereas 8% and 4% intend to invest more than € 10 monthly for mobile applications and digital music respectively.

We investigated the ways people prefer to explore new music and came to the clear result that for the majority of respondents television is not an important source with which to discover music. In contrary, internet and friends are used as sources to explore music by the most people in the sample, this is displayed in Figure 4.3. It is assumed that the combination of these two sources in social communities might have even larger spread.

Music is considered to be perceived as a private matter, according to the results of the survey, respondents are generally more willing to explore music than to let someone else

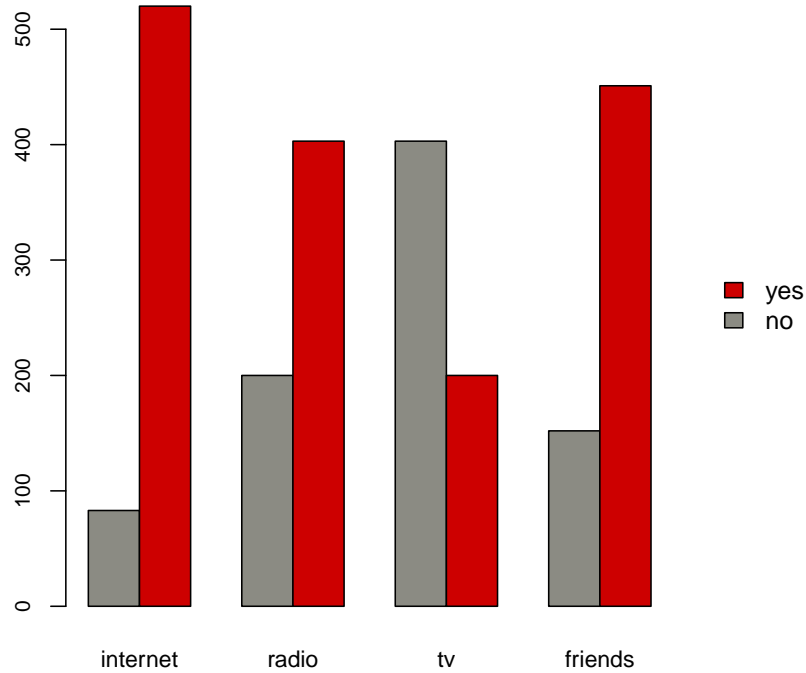


Figure 4.3: Preference for different sources of music exploration

explore own music tastes. This is true for all levels of familiarity with the person, except of friends, where the rates are approximately the same.

The most prominent difference is seen with the group of professionals, where 43% are willing to listen, but only 21% are willing to share music, which is illustrated in Figure 4.5. This tendency can be explained by the fact that common music listeners are consumers of music and do feel negative about sharing their music tastes with people, who are professional in the music industry. With decreasing level of familiarity, from friends to unknown people, the number of people wishing to listen into or to share music declines. The most obvious difference can be observed by comparing willingness to listen or share music in the group of friends and the group of unknown people. Whereas the proportion of people willing to listen/share music with friends does not differ and equals approximately 86%, see Figure 4.4, willingness to listen to music of unknown people is slightly higher than to share music with unknown people (23% and 18% respectively), which is three times less than in the group of friends, see Figure 4.6.

Finally, we compare the proportion of people, who stated their positive WTP in the hypothetical WTP question, with people, who gave a "definitely sure" response to the

4 Design of empirical study

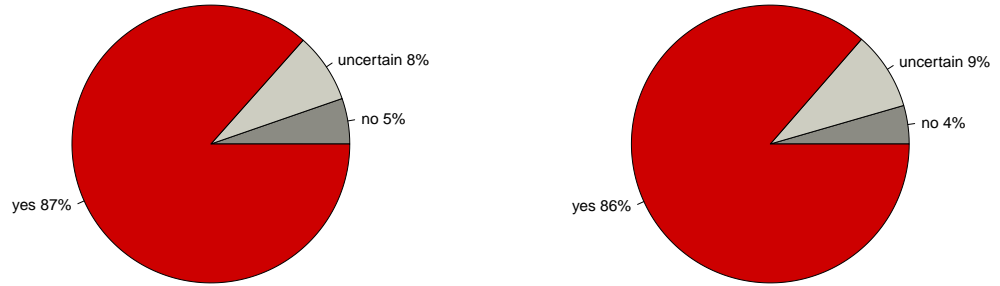


Figure 4.4: Willingness to listen (left) and share (right) music with friends

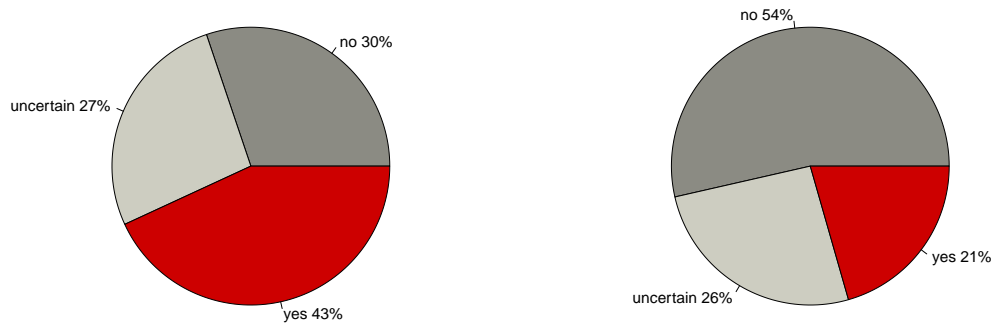


Figure 4.5: Willingness to listen (left) and share (right) music with professionals

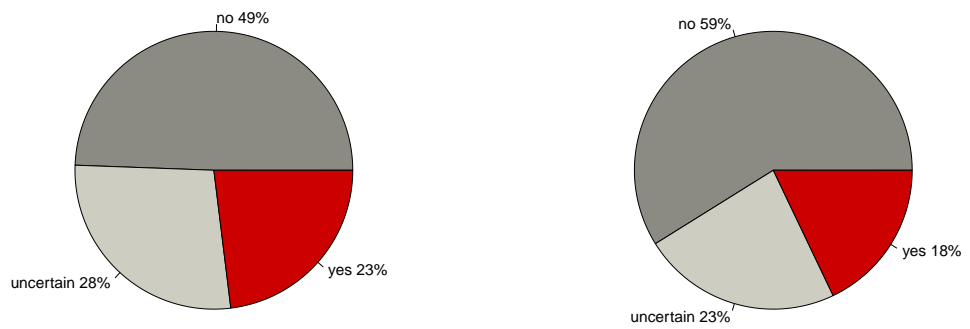


Figure 4.6: Willingness to listen (left) and share (right) music with unknown people

4 Design of empirical study

calibrating certainty question. From Figure 4.7 we can conclude that only a considerably t number of people supported their hypothetical decision with "definitely sure" statement. Hence, from the originally observed percentages of positive statements 26.4%, 10.3%, 21.6% and 12.6%, after calibration 5.6%, 3.4%, 5.8% and 5.1% are expected for the four virtual goods respectively. Taking into account the *Zynga's* assessment, Reuters (2011) and *eMarket* experts' evaluation of the virtual goods market, the virtual goods monetization level lies between 2 – 6%. Therefore, the calibrated values are closer to the market benchmark while hypothetical values are highly overestimated.

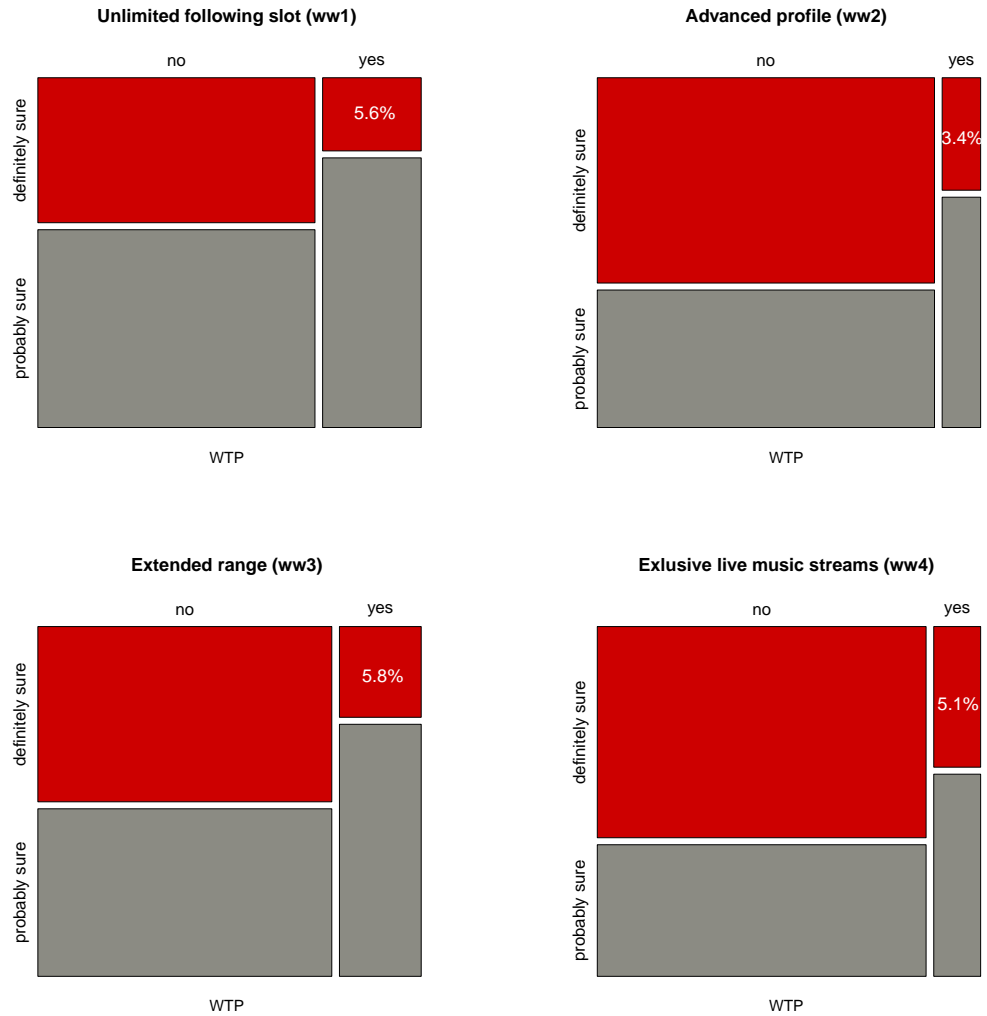


Figure 4.7: Hypothetical WTP versus calibrated WTP response rates

5 Willingness to pay prediction with logistic regression

5.1 Logistic regression model

Multiple logistic regression, also called a logit model, describes the relationship between a dichotomous response variable Y and multiple explanatory variables denoted by X representing the whole set of covariates x_1, \dots, x_p , which can be either continuous or categorical. The dependent variable Y is binary or dichotomous and can take values of 0 and 1 for non-purchase and purchase, respectively, Hosmer and Lemeshow (1989).

The conditional mean represents the expected value of the response variable Y , given the value of the independent variable x is denoted as $P(Y|x)$. In linear regression it is possible for $P(Y|x)$ to take any values $(-\infty; \infty)$, but with dichotomous response variable the conditional mean is bounded between 0 and 1, i.e. $[0 \leq P(Y|x) \leq 1]$.

For simplification purposes the conditional mean $P(Y|x)$ is further denoted as $\pi(x)$ at each value of x 's and $\pi(x)$ is calculated as:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}} \quad (5.1)$$

The logit transformation of $\pi(x)$ is defined in terms of $\pi(x)$ as:

$$g(x) = \text{logit} \{ \pi(x) \} = \ln \left\{ \frac{\pi(x)}{1 - \pi(x)} \right\} \quad (5.2)$$

Systematic component of the multiple logistic regression is a linear predictor with more than 1 variable $\alpha + \beta_1 x_1 + \dots + \beta_k x_k$. For the *logit* of $\pi(x)$ logistic regression model has linear form:

$$g(x) = \text{logit} \{ \pi(x) \} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (5.3)$$

In our study the explanatory variables are either dichotomous or categorical with $k_j \geq 2$ levels, where k_j is number of categories of the j^{th} independent variable. We can represent the *logit* $\{ \pi(x) \}$ in terms of design variables, where $k_j - 1$ design variables are needed to estimate the model, Hosmer and Lemeshow (1989). Design variables can be denoted as D_{jm} , where m signifies the levels of independent variable, $m = 1, 2, \dots, k_j - 1$ and j

stands for the j^{th} independent variable.

The equation of the multiple logistic regression given in terms of design variables is given below.

$$g(x) = \text{logit} \{ \pi(x) \} = \beta_0 + \sum_{j=1}^p \sum_{m=1}^{k_j-1} \beta_{jm} D_{jm},$$

where β_{jm} denotes the coefficient of design variable D_{jm} .

Hence, the $\pi(x)$ can be denoted as:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}. \quad (5.4)$$

5.2 Fitting the logistic regression model

To estimate the regression parameters, logistic regression employs the maximum likelihood estimation (MLE) method. According to Hosmer and Lemeshow (1989), the idea of the MLE can be described as searching for parameters that maximise the probability of obtaining the observed data. At the first step the likelihood function should be constructed, which expresses the probability of the observed data as a function of the unknown parameters.

At the second step, the maximum likelihood estimators of these parameters are chosen to maximise the likelihood function. In the multivariate case, β' is the vector of parameters, i.e. $\beta' = (\beta_0, \beta_1, \dots, \beta_k)^T$. The conditional probability of purchase $Y = 1$, given x is denoted as $P(Y = 1|x) = \pi(x)$, whereas probability of no-purchase $P(Y = 0|x) = 1 - \pi(x)$.

Therefore, for the sample of n independent observations, for the pairs (x_i, y_i) , where x_i is the value of the independent variable and y_i is the value of the dependent variable for the i^{th} subject, the contribution to the likelihood function, when $y_i = 1$ and $y_i = 0$ are $\pi(x_i)$ and $1 - \pi(x_i)$ respectively. Hence, the contribution of the pair (x_i, y_i) to the likelihood function can be calculated as:

$$\zeta(x_i) = \pi(x_i)^{y_i} \{1 - \pi(x_i)\}^{1-y_i} \quad (5.5)$$

While the independence of the observations is assumed, the likelihood function for the n observations is given as:

$$l(\beta) = \prod_{i=1}^n \zeta(x_i) = \prod_{i=1}^n \left[\pi(x_i)^{y_i} \{1 - \pi(x_i)\}^{1-y_i} \right] \quad (5.6)$$

It is easier to work with log likelihood function, which is given as follows:

$$\log \{l(\beta)\} = \sum_{i=1}^n [y_i \log \{\pi(x_i)\} + (1 - y_i) \log \{1 - \pi(x_i)\}] \quad (5.7)$$

$\hat{\beta}$ is the maximum likelihood estimator of β , $\hat{\pi}(x_i)$ is the maximum likelihood estimate of $\pi(x_i)$ computed using $\hat{\beta}$ and x_i .

5.3 Interpretation of the logistic regression parameters

Independent variables in our data set are dichotomous ($k_j = 2$) or categorical ($k_j > 2$). In this section using the variable *age*, which has four levels, we provide the interpretation of the regression coefficients, whereas dichotomous variables are considered as a sub-case of the categorical independent variables.

First, it is necessary to build a set of design variables, which represent the categories of the variable *age*, $k_1 = 4$, where $j = 1$ for *age* variable, though we need $k_1 - 1 = 3$ design variables. We use the "18 – 24" as a reference group and the specification of the design variables is provided in Table 5.1.

The method for specifying the design variables we employ requires setting all of them to zero for the reference group and then setting each of a single design variable to 1 for each of the other groups as in Hosmer and Lemeshow (1989).

age	Design variables		
	D_{11}	D_{12}	D_{13}
18-24 (1)	0	0	0
25-31 (2)	1	0	0
32-38 (3)	0	1	0
39-45 (4)	0	0	1

Table 5.1: Design variables

The probability of success for every cell for the age groups "18 – 24" and "25 – 31", where n_{11}, \dots, n_{24} represent the number of observations corresponding to the respective cases, is calculated as following:

1. $\phi_{11} = n_{11}/n_{1\bullet}$;
2. $\phi_{21} = n_{21}/n_{2\bullet}$;
3. $\phi_{12} = n_{12}/n_{1\bullet}$;
4. $\phi_{22} = n_{22}/n_{2\bullet}$.

Table 5.2 provides the cross-classification of the levels of *age* variable and response variable,

age	purchase	non-purchase	total
18-24	n_{11}	n_{12}	$n_{1\bullet}$
25-31	n_{21}	n_{22}	$n_{2\bullet}$
32-38	n_{31}	n_{32}	$n_{3\bullet}$
39-45	n_{41}	n_{42}	$n_{4\bullet}$
total	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Table 5.2: Cross-classification table

Now it is possible to derive the odds ratios ($\hat{\psi}$). For instance, let us calculate the $\hat{\psi}$ for the age group "25 – 31" with reference group "18 – 24":

$$\hat{\psi}(\text{"25 – 31"}, \text{"18 – 24"}) = \frac{\phi_{11}\phi_{22}}{\phi_{12}\phi_{21}} \quad (5.8)$$

Furthermore, $\log \left\{ \hat{\psi}(\text{"25 – 31"}, \text{"18 – 24"}) \right\} = \hat{\beta}_{11}$, which can be derived from the following equations.

To compare the age group "25 – 31" with "18 – 24", we have to calculate the estimate of the log odds, which is the difference between estimated logits computed at two levels. The estimated logit of the group "18 – 24" is equal to:

$$g(\text{"18 – 24"}) = \left[\hat{\beta}_0 + \hat{\beta}_{11}(D_{11} = 0) + \hat{\beta}_{12}(D_{12} = 0) + \hat{\beta}_{13}(D_{13} = 0) \right], \quad (5.9)$$

whereas the estimated logit of the group "25 – 31" is calculated as:

$$g(\text{"25 – 31"}) = \left[\hat{\beta}_0 + \hat{\beta}_{11}(D_{11} = 1) + \hat{\beta}_{12}(D_{12} = 0) + \hat{\beta}_{13}(D_{13} = 0) \right] \quad (5.10)$$

The logit difference is:

$$\log \left[\hat{\psi}(\text{"25 – 31"}, \text{"18 – 24"}) \right] = \hat{g}(\text{"25 – 31"}) - \hat{g}(\text{"18 – 24"}) = \hat{\beta}_{11} \quad (5.11)$$

5.4 Model selection

To select the best model we employ the backward stepwise variable selection procedure. This algorithm begins with a model, which contains all predictor variables and at each stage removes the variable with the largest *p*-value in the test so that its parameters equal zero. The algorithm will stop deletion when deletion of any further variable leads to a significantly poorer fit.

Akaike information criterion (AIC) measures the goodness of fit and will be calculated for every stage of elimination. The optimal model minimises:

$$AIC = -2 \log \{l(\beta)\} + 2p \quad (5.12)$$

and has its fitted values closest to the true outcome probabilities. $-2 \log \{l(\beta)\}$ is a badness-of-fit indicator, that is, large values mean poor fit of the model to the data. p is the number of estimated parameters.

After selecting the model with the lowest AIC, we run the analysis of deviance to compare two models, a null model with intercept only and a model containing covariates in order to distinguish how well the chosen logit model fits the data. The difference between the maximised value of the likelihood functions for the null model l_0 and a full model l_1 should be calculated. L_0 and L_1 denote the maximised log-likelihood functions. The formula for the likelihood-ratio test statistic G is:

$$G = -2 \log \left(\frac{l_0}{l_1} \right) = -\{2 \log(l_0) - 2 \log(l_1)\} = -2(L_0 - L_1), \quad (5.13)$$

while for large samples G is χ^2 distributed.

The model with covariate(s) fits better in comparison to the null model, when the test statistic G is large with respectively small p -values.

5.5 Empirical results

We ran the logistic regression for the four response variables in our data set. Next we employ function `step` for the backward variables selection procedure in order to select the best model with the smallest AIC measure.

The variables in the logistic regression models were design variables of the categorical variables, with first category of each variable taken to be reference group, see Table 5.3. The results of the four best logistic models are given in the Tables 5.4 - 5.7.

The antilog of a $\hat{\beta}$ parameter estimate in logistic regression is a multiplicative effect on the odds for the response variable, for each one level increase in the predictor (design) variable of which it is a coefficient. Hence, for logistic regression the odds ratio is a common measure of the nature and strength of an association between independent and dependent variables.

Considering the first response variable, with other variables being fixed, the probability of purchase decreases with age, increases with a positive music budget, willingness to listen to professionals and to share music with social networks as well as unknown people, and past experience with buying the virtual goods.

Age has a significant influence on the purchase probability so that latter decreases from

initial variable	design variable	category
gen	gen	female
	age1	25-31
age	age2	32-38
	age3	39-45
	bapp1	< 5 €
bapp	bapp2	5 - 10 €
	bapp3	> 10 €
	bmus1	< 5 €
bmus	bmus2	5 - 10 €
	bmus3	> 10 €
	tv	tv
lfr	lfr1	uncertain
	lfr2	yes
lac	lac1	uncertain
	lac2	yes
lsc	lsc1	uncertain
	lsc2	yes
lpr	lpr1	uncertain
	lpr2	yes
lun	lun1	uncertain
	lun2	yes
sfr	sfr1	uncertain
	sfr2	yes
sac	sac1	uncertain
	sac2	yes
ssc	ssc1	uncertain
	ssc2	yes
spr	spr1	uncertain
	spr2	yes
sun	sun1	uncertain
	sun2	yes
soga	soga1	uncertain
	soga2	yes
enlev	lev1	low
	lev2	middle
	lev3	high
vigo	vigo	yes

Table 5.3: Assignment of design variables

young to old. Being in the age group "25 – 31" versus age group "18 – 24" decreases the chances of purchase by $\exp\{-0.5911\} = 0.55$ times; that is by 45%. Having a low monthly budget for music increases the probability of purchase by 2.4 times. Willingness

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.2177	0.3118	-7.113	1.13e-12	***
gen	0.3724	0.2293	1.624	0.104377	
age1	-0.5911	0.2432	-2.431	0.015060	*
age2	-0.6845	0.3571	-1.917	0.055272	.
bapp3	-0.7252	0.4398	-1.649	0.099219	.
bmus1	0.8664	0.2379	3.642	0.000271	***
bmus2	0.4963	0.3181	1.560	0.118728	
tv	-0.6197	0.3106	-1.995	0.046045	*
lsc1	-0.4054	0.2359	-1.718	0.085734	.
lpr1	0.4996	0.3017	1.656	0.097730	.
lpr2	0.5665	0.2746	2.063	0.039122	*
sfr1	0.7041	0.3779	1.863	0.062477	.
ssc1	0.5905	0.2729	2.164	0.030474	*
ssc2	0.5909	0.2915	2.027	0.042677	*
spr1	0.3850	0.2498	1.541	0.123203	
sun2	0.4901	0.2942	1.666	0.095808	.
soga2	-0.9420	0.5633	-1.672	0.094450	.
lev2	0.4930	0.3301	1.493	0.135346	
vigo	0.8400	0.3307	2.540	0.011084	*
Null deviance:	639.05	on 560 df			
Residual deviance:	570.66	on 542 df			
AIC:	608.66				

Table 5.4: Best logistic model for the unlimited following slot (ww1), with ***, **, * and . corresponding to significance levels of 0.001, 0.01, 0.05, 0.1 respectively and non-significant variables marked grey

to listen to the music of professionals has the effect of multiplying the estimated odds of purchase by 1.76. Sharing music with people from social networks has a strong positive influence on the estimated odds of purchase of 1.8. Past experience of purchasing virtual goods increases the probability of purchase by 2.3 times in comparison to people, who have never bought virtual goods.

Similarly in the second model, age has a negative influence and a monthly budget for music a positive influence on purchase probability. Conversely, willingness to listen to professionals decreases the purchase probability by 0.4 times. Willingness to listen to unknown people has a positive influence on the purchase probability, whereas the high engagement level with social games increases the odds of purchase by 4.7 times.

Also, for the third feature the fact of spending money for music has a strong significant influence on the odds of purchase, increasing it 2.7 times. Using internet as a source of music, increases the probability of purchase. Willingness to share music with social

5 Willingness to pay prediction with logistic regression

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-2.6476	0.2782	-9.516	< 2e-16	***
age1	-1.0453	0.3932	-2.659	0.007847	**
bmus1	0.8401	0.3798	2.212	0.026974	*
bmus2	0.9950	0.3873	2.569	0.010201	*
lfr1	1.4604	0.4436	3.292	0.000995	***
lpr2	-0.8780	0.3550	-2.473	0.013403	*
lun2	0.7574	0.3519	2.153	0.031357	*
sun1	0.5474	0.3432	1.595	0.110717	
lev3	1.5564	0.6006	2.592	0.009555	**
vigo	0.6904	0.4302	1.605	0.108536	
Null deviance:	355.43	on 560 df			
Residual deviance:	311.34	on 550 df			
AIC:	333.34				

Table 5.5: Best logistic model for the advanced profile (ww2), with ***, **, * and . corresponding to significance levels of 0.001, 0.01, 0.05, 0.1 respectively and non-significant variables marked grey

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-3.9006	0.6262	-6.229	4.69e-10	***
bmus1	1.0022	0.2478	4.044	5.26e-05	***
bmus2	0.7821	0.3219	2.430	0.015111	*
int	1.0554	0.4339	2.432	0.014999	*
lfr1	-1.0538	0.6346	-1.660	0.096838	.
lsc1	-0.4688	0.2537	-1.848	0.064643	.
lun1	0.6365	0.2387	2.666	0.007677	**
sfr2	0.8366	0.4696	1.782	0.074819	.
ssc1	0.6154	0.2786	2.209	0.027158	*
ssc2	0.6036	0.2802	2.154	0.031235	*
vigo	1.1508	0.3320	3.466	0.000528	***
Null deviance:	585.00	on 560 df			
Residual deviance:	511.94	on 550 df			
AIC:	533.94				

Table 5.6: Best logistic model for the extended range (ww3), with ***, **, * and . corresponding to significance levels of 0.001, 0.01, 0.05, 0.1 respectively and non-significant variables marked grey

networks also has a significant effect on the odds of purchase multiplying it 1.8 times. Similarly to the first model, the past experience of purchasing virtual goods has the strongest significant positive influence.

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-4.5733	0.8747	-5.228	1.71e-07	***
gen	0.5734	0.3065	1.871	0.06138	.
bmus1	0.9388	0.3200	2.934	0.00335	**
bmus2	1.3246	0.4092	3.237	0.00121	**
bmus3	1.7640	0.5653	3.120	0.00181	**
int	1.0033	0.5631	1.782	0.07478	.
fr	-0.8735	0.2933	-2.979	0.00290	**
lsc1	-0.6655	0.3208	-2.074	0.03805	*
lpr2	0.5268	0.2801	1.881	0.05997	.
sfr2	1.0318	0.6570	1.571	0.11626	
sac2	0.6086	0.2903	2.096	0.03607	*
ssc1	1.0198	0.3160	3.227	0.00125	**
sun1	-0.5779	0.3639	-1.588	0.11227	
soga1	-1.1963	0.3955	-3.025	0.00249	**
lev1	0.7208	0.4130	1.745	0.08091	.
lev2	0.9081	0.4944	1.837	0.06623	.
Null deviance:	426.13	on 560 df			
Residual deviance:	358.76	on 545 df			
AIC:	390.76				

Table 5.7: Best logistic model for the exclusive live music streams (ww4), with ***, **, * and . corresponding to significance levels of 0.001, 0.01, 0.05, 0.1 respectively and non-significant variables marked grey

In the fourth model a monthly budget for music also increases the probability of purchase, while using friends as a source of music has a negative effect.

Eventually we perform an ANOVA analysis of deviance, a likelihood-ratio test (LRT) is computed as the difference between deviance of the full model and model with intercept only.

The model with covariates fits better in comparison to the null model, when the test statistic is large with respectively small p -values. Single covariates are added to the null model sequentially from the first to the last.

The resulting likelihood-ratio test statistic is χ^2 distributed, with degrees of freedom equal to the number of parameters that are constrained. The associated p -values, which are $p < 0.001$, indicate that the models with selected predictors fit significantly better than the model with only an intercept, see Table 5.8.

5 Willingness to pay prediction with logistic regression

Model	Null deviance	Res. deviance	LRT	df	p-value
ww1	639.05	570.66	68.39	18	8.4433e-08
ww2	355.43	311.34	44.09	10	3.1713e-06
ww3	585.00	511.94	73.06	10	1.1335e-11
ww4	426.13	358.76	67.37	15	1.3079e-08

Table 5.8: Likelihood-ratio test results

Summarising the logistic regression results of the four models, we can conclude that one factor that is the most decisive for all four models is the monthly budget for music, whereas the fact of spending money for music versus not spending money is critical. Furthermore, for the first and the second models with increasing age the probability of purchase declines and willingness to listen to professionals as well as willingness to share music with social networks increase the odds of purchase. Moreover, the first and third models share other significant variables, e.g. past experience with buying virtual goods.

6 Willingness to pay prediction with CART

CART - Classification and Regression Tree is a non parametric method that employs available data from the past and tries to explain a relationship between explanatory and exploratory variables in the form of a binary tree, developed by Breiman et al. (1984). If a response variable is categorical, CART produces a classification tree, otherwise if a response variable is represented by a continuous variable, a regression tree is produced. When new observations are available, it is possible to classify them according to classes of exploratory variable, by the means of the constructed decision tree.

Tree-based methods have been widely used in computer sciences, health care, ecology and decision making in financial markets. However, application of CART in marketing decision making is not prevalent. CART attracts researchers due to its ease of interpretation and understanding as a series of *if-then* relationships.

In this chapter we intend to introduce the basic principles of CART methodology. In addition, we aim to illustrate the effectiveness of CART in comparison to logistic regression.

The purpose of CART is to provide such classification rules that enable the prediction of the class (purchase / no purchase) of any further observations, given the set of characteristics submitted for analysis. The probability of occurrence is assigned to each end of branch in the tree, Timofeev (2010).

CART splits a sample into binary sub samples (left and right nodes) based on the response to a dichotomous question with yes/no answer, based only on a single variable. There are two types of nodes: nodes which do not split further are called terminal nodes, whereas those which have further splits are non-terminal nodes.

The purpose of building a CART tree is to:

- determine the optimal splitting rule with best split s^* at each node
- determine the optimal tree size T^*
- apply the T^* to classify new data

6.1 Growing the classification tree

Let N be the number of observations in our sample and N_j - the number of observations of class j , $j = \overline{1, J}$. We can then define the distribution of the classes $\pi(j)$ as the

proportion of the classes in the population:

$$\pi(j) = \frac{N_j}{N}, \quad (6.1)$$

for $j = \overline{1, J}$.

Analogically $N(t)$ is the number of observations in node t and $N_j(t)$ is the number of observations of class j in the node t . Below we can define the joint probability of an observation of j -th class to fall into node t as:

$$p(j, t) = \pi(j) \times \frac{N_j(t)}{N_j}, \quad (6.2)$$

so we can derive that $p(t) = \sum_{j=1}^J p(j, t)$. The conditional probability of an observation of node t given its class is j is calculated as:

$$p(j|t) = \frac{p(j, t)}{p(t)} = \frac{N_j(t)}{N(t)}, \quad (6.3)$$

in other words $p(j|t)$ is class probability distribution at node t , whereas $\sum_{j=1}^J p(j|t) = 1$. At this stage we are interested in finding the optimal split s^* at the node t , for which class homogeneity for a given tree node is the highest. Class homogeneity is defined by impurity function $\phi(t)$ and impurity measure $i(t)$. Impurity function $\phi(t)$ is a function of class probabilities $p(1|t), p(2|t), \dots, p(J|t)$ and is determined on subsets $\{p_1, p_2, \dots, p_J\}$ for any J and $p_j \geq 0, j = \overline{1, J}, \sum_{j=1}^J p_j = 1$. The unique maximum of the impurity function is attained, when all classes in the population have equal probability of occurrence: $p(1|t) = p(2|t) = \dots = p(J|t)$. The unique minimum of the impurity function ($\phi(t) = 0$) is achieved, when all classes of the node belong to one class: $p(J|t) = 1$. So, given the impurity function ϕ , we can define the impurity measure $i(t)$ for the node t as:

$$i(t) = \phi [p(1|t), p(2|t), \dots, p(J|t)] \quad (6.4)$$

It is now possible to derive the goodness-of-split criteria of the split s at a node t . Though, for the parent node t , there are two child nodes, i.e. t_L and t_R representing the left and right nodes respectively. Then the goodness-of-split criteria can be measured as the reduction in impurity at the node t :

$$\Delta i(s, t) = i(t) - p_L i(t_L) - p_R i(t_R), \quad (6.5)$$

where p_L, p_R is the fraction of cases at node t that fall into the t_L and t_R and $i(t_L), i(t_R)$ is an impurity measure of the t_L and t_R respectively. At each node the following

optimisation problem is solved:

$$\begin{aligned}
s^* &= \arg \max_s \Delta i(s, t) \\
&= \arg \max_s \{-p_L i(t_L) - p_R i(t_R)\} \\
&= \arg \min_s \{p_L i(t_L) + p_R i(t_R)\}
\end{aligned} \tag{6.6}$$

CART selects the best split s^* of the variable, for which the reduction in impurity is maximised, in other words, for which the homogeneity of the child nodes is the highest. The different splitting rules exist, two of the most commonly used are the *Twoing splitting rule* and the *Gini criterion*. Twoing splitting rule sorts out two classes, which result in more than 50% of the data. Since our explanatory variable has only two classes, the *Gini splitting criterion* is preferred. According to the Gini criterion the largest class in a learning sample is separated from the rest of the data. The impurity function employing the *Gini criterion* can be defined as:

$$i(t)_{gini} = 1 - \sum_{j=1}^J p^2(j|t) \tag{6.7}$$

The *Gini criterion* is derived from the sample variance estimate at node t over all classes of the dependent variable:

$$\begin{aligned}
\sum_{j=1}^J [p(j|t) \{1 - p(j|t)\}] &= \sum_{j=1}^J \{p(j|t) - p(j|t)^2\} \\
&= \sum_{j=1}^J p(j|t) - \sum_{j=1}^J p(j|t)^2 \\
&= 1 - \sum_{j=1}^J p^2(j|t)
\end{aligned}$$

where $\sum_{j=1}^J p(j|t) = 1$.

6.2 Tree pruning methods

One of the most important questions of the tree construction is defining the optimal size of the tree, in order to avoid either underspecification or overspecification of the parameters. Overspecification problems often occur in the case of a *maximum tree*. Maximum tree splits learning sample into absolutely class homogeneous groups and though has low or zero misclassification rates. However it is difficult to interpret trees with large numbers of terminal nodes. Moreover, since a maximum tree considers any small and insignificant variations, it provides poor results when applied to new datasets.

Underspecification is a problem of too small trees, where only a few iterations were used to split the dataset. In such a constellation significant relationships probably could not be revealed. In this section two methods of tree pruning, such as cross-validation and cost-complexity function, are described. Pruning of the tree can be described as collapsing some of the branches of the T_{max} from the bottom up.

The aim of the cross-validation procedure is to extract maximum information from the learning sample, so that the available data is employed alternating as a training or as a test sample, while the larger proportion of observations are assigned to training set and the rest is used to verify the tree quality. Such a procedure is possible since the actual class value of the dependent variable is available from the learning sample.

The learning sample is randomly divided into K parts, whereas the training set is denoted as $(K - 1)$ and $\frac{1}{K}$ stands for the test set. In the next step the data used previously as the test sample becomes a part of the training sample and the other $\frac{1}{K}$ becomes a test sample. The procedure is continuous until all the data points are employed both as training and test samples. For a given classification rule $d^{(k)}$ and for training sample $K - 1$ and since none of the observations of the test set was involved in the construction of the classification rule $d^{(k)}$, it is possible to define the cross-validation measure of tree quality as:

$$E^{CV}(d) = \frac{1}{K} \sum_{k=1}^{K-1} E^1(d^{(k)}), \quad (6.8)$$

where $E^1(d^{(k)})$ is a one-iteration estimate.

It is suggested that cross-validation procedure with $K = 10$ provides an acceptable level of result robustness. 10-fold cross-validation means that the following procedure is repeated 10 times: a 10% random sample is selected from the learning sample, the model is fitted to the remaining 90%, and a prediction is made from the fitted model for the selected 10%.

A cost-complexity method takes into account the trade-off between accuracy and complexity of the tree. While the complexity is defined by the number of terminal nodes, the relationship between accuracy and complexity is as following: the smaller the tree, the more limited prediction power it has, but it is less complex. Whereas a maximum tree can provide perfect in-sample predictions, but it obtains a complexity penalty because of its large size.

For any subtree $T < T_{max}$, the number of terminal nodes is denoted as $|\tilde{T}|$. The following cost-complexity function is employed to optimise classification tree size:

$$E_\alpha(T) = E(T) + \alpha |\tilde{T}|, \quad (6.9)$$

where $\alpha \geq 0$ represents the complexity penalty for additional terminal node. $\alpha |\tilde{T}|$ is

a cost parameter. $E(T)$ is an internal misclassification tree error, defined as the sum of internal misclassification errors at every node t , $t \in \tilde{T}$. The higher the number of terminal nodes, the lower the misclassification tree error is, but the higher the complexity of the tree is and vice versa.

Further we search for any $\alpha \geq 0$ the optimal tree $T(\alpha)$, that minimizes the $E_\alpha(T)$, Breiman et al. (1984):

$$E_\alpha \{T(\alpha)\} = \min_{T \leq T_{max}} E_\alpha(T) \quad (6.10)$$

The $T(\alpha)$ are pruned trees of the maximum tree T_{max} . For $\alpha = 0$, we denote the pruned subtree as T_1 .

Further procedure of tree pruning is as follows: T_1 is found, weak link \bar{t}_1 is detected and branch $T_{\bar{t}_1}$ is pruned off, then α_2 is calculated and the process is continued. In such way the new tree $T_2 \prec T_1$ is defined by:

$$T_2 = T_1 - T_{\bar{t}_1} \quad (6.11)$$

With growing α the tree will be shorter until the root node $T \{0\}$ is reached. Although α is infinite, the number of pruned subtrees which minimise $E_\alpha(T)$ is finite:

$$T_{MAX} \succ T_1 \succ T_2 \succ \dots \succ T \{0\} \quad (6.12)$$

Now, by applying the method of K-fold cross-validation to the tree sequence given in the equation 6.8 the optimal tree can be determined.

However, selecting a tree with the minimum value of $E^{CV}(T)$ is not appropriate, because usually the whole range of $E^{CV}(T)$ which satisfy $E^{CV}(T) < E_{min}^{CV}(T) + \varepsilon$ for small $\varepsilon > 0$ exists. If $K < N$ then the second run of the cross-validation procedure will provide different results. Therefore, it is suggested rather to apply one standard error empirical rule, according to which if T_{k_0} is the tree minimising $E^{CV}(T_{k_0})$ from the sequence of the equation 6.12, then the value k_1 and the corresponding tree T_{k_1} are selected so that:

$$\arg \max_{k_1} \hat{E}(T_{k_1}) \leq \hat{E}(T_{k_0}) + \sigma \left\{ \hat{E}(T_{k_0}) \right\}, \quad (6.13)$$

where $\sigma(\cdot)$ is the sample standard error estimate and $\hat{E}(\cdot)$ stand for the internal misclassification errors estimates.

6.3 Empirical results

The input vector of explanatory variables $X = (X_1, X_2, \dots, X_p)$ contains features of categorical variables. For categorical variables $X_j \in \{1, 2, \dots, M\}$ there is a set Q of

binary splits s^* in form of the question:

$$\{is\ X_j \in A\}, \quad (6.14)$$

where A is a independent variable value, which ranges over all subsets of $\{1, 2, \dots, M\}$. Left nodes stand for the positive answers, right nodes for the negative ones. Split s^* data sample is divided into two sub-samples, so that homogeneity within each sub-sample is ensured.

Every question in a tree splits the initial data into two parts, so that the splitting procedure is repeated until the optimal tree T^* is reached and the binary splits constitute the standard set of questions.

In the current study, initial maximum trees for the four dependent variables were grown using most significant predictor variables listed in the Table 6.1 from the set of the independent variables described in the Table 4.3. After the initial classification trees had been grown, the trees were subsequently pruned.

A cost-complexity method was used on the learning sample to determine the optimal number of nodes of the tree, so that the relationship between accuracy and complexity was optimized for the tree. Predictions were obtained on the test data set using the pruned tree. The classification tree models were fit using the `tree()` function of the `tree` package in R.

On the following Figure 6.1 we represent the cost-complexity pruning tree sequence. On the basis of these graphics, we can choose the optimal classification tree, considering the trade-off between accuracy, i.e. misclassification tree error and complexity, i.e. number of terminal nodes $|\tilde{T}|$ of the tree.

Starting with the first model, we can conclude that misclassification error reduction effectiveness clearly decreases when the size of the tree reaches 18 terminal nodes. The misclassification error lies by the minimum of 70 for the maximum tree, which correspond to the 15.2% misclassification error. For this reason the optimal pruning point lies at 18 terminal nodes, providing only a slightly higher misclassification error of 19.4%.

The classes of the variable $ww2$ are classified much better in comparison to the classification of $ww1$. The maximum tree has the total of 43 terminal nodes and misclassification error rate of 8.0%. With the pruning of the tree to 14 terminal nodes, the misclassification rate did not increase.

Rather analogically to the first variable $ww1$, the maximum tree for the variable $ww3$ also has a very large number of terminal nodes 68, which makes the interpretation almost impossible. The number of misclassifications does not lie under the 79 mark, which corresponds to the 14.1% misclassification rate. The misclassification rate has increased but not substantially and equals 16.2% with the pruning of the tree to the size of 15 terminal nodes.

6 Willingness to pay prediction with CART

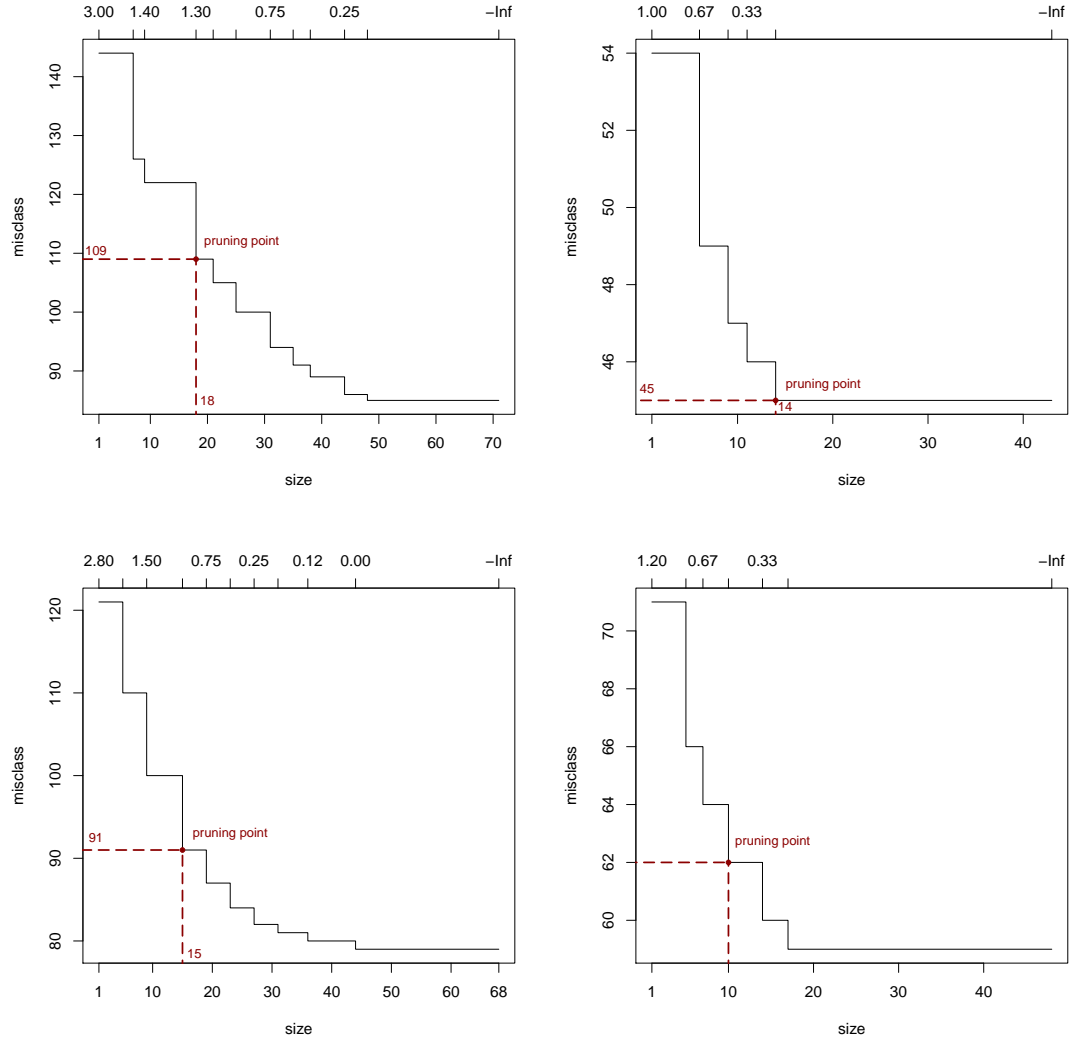


Figure 6.1: Cost-complexity pruning tree sequence statistics for the four dependent variables *ww1* (top left), *ww2* (top right), *ww3* (bottom left), *ww4* (bottom right), showing the number of terminal nodes in each tree in the sequence, the total number of misclassifications of each tree and accordingly to each tree - the value of the cost-complexity pruning parameter.

The maximum tree for the variable *ww4* contains 48 terminal nodes and its misclassification rate is 10.5% or 59 misclassified cases. The pruning procedure shows that with 10 terminal nodes, the number of misclassifications is 62, which is correspondent to less than 1% loss compared to the maximum tree. While the misclassification rate is 11.1%. CART sorts out the significant variables, which are then used in order to build a tree. These are given in Table 6.1 both for the maximum tree and for the pruned tree, in the order they were used in the tree construction process, starting with the root node. It is obvious that the independent variable *monthly budget for music* is the most important

6 Willingness to pay prediction with CART

Var	Maximum tree	N	Reduced tree	N
ww1	bmus, ssc, sogalpr, lsc, sun, int, sac, lfr, fr, rad, bapp, enlev, age, lun, lac, os, ViGo, spr, gen, tv	21	bmus, ssc, sogalpr, sun, age, lun, lac, rad, sac, ViGo, spr, gen, tv	13
ww2	bmus, lfr, enlev, age, lac, sac, fr, bapp, lpr, ssc, sfr, sun, ViGo, int, os, gen, spr, lun, os, sogalpr	20	bmus, lfr, enlev, age, bapp, sun, lun, spr, os, ssc, lpr	11
ww3	bmus, ViGo, lac, int, enlev, ssc, age, fr, sogalpr, bapp, lfr, spr, sac, lpr, gen, sun, os, lsc, lun, sfr, rad	21	bmus, lun, enlev, age, lpr, gen, bapp, sac, lac, spr, lsc	11
ww4	bmus, fr, sun, sac, tv, enlev, lsc, ssc, rad, lun, lpr, gen, bapp, age, os, spr, lac, sogalpr	18	bmus, spr, lac, ssc, bapp, lpr sac, lun	8

Table 6.1: Significant variables used in the construction of the maximum and reduced classification tree

variable in all four classification models.

In particular classification tree of *ww1*, but also of *ww3* both provide considerably high misclassification rates 19.4% and 14.1% respectively. This means that the prediction of responses for other people which are not included in our data sample, may be poor. The second tree of *ww2* has an 8.2% misclassification rate and a fourth tree - 10.2%. This is still a quite high accuracy loss.

Due to considerably high misclassification rates in the pruned trees, the accuracy loss appears to be high. For this reason, in the next chapter, we perform the model assessment analysis of in-sample and out-of-sample settings.

Figures 6.2 - 6.5 depict pruned classification trees.

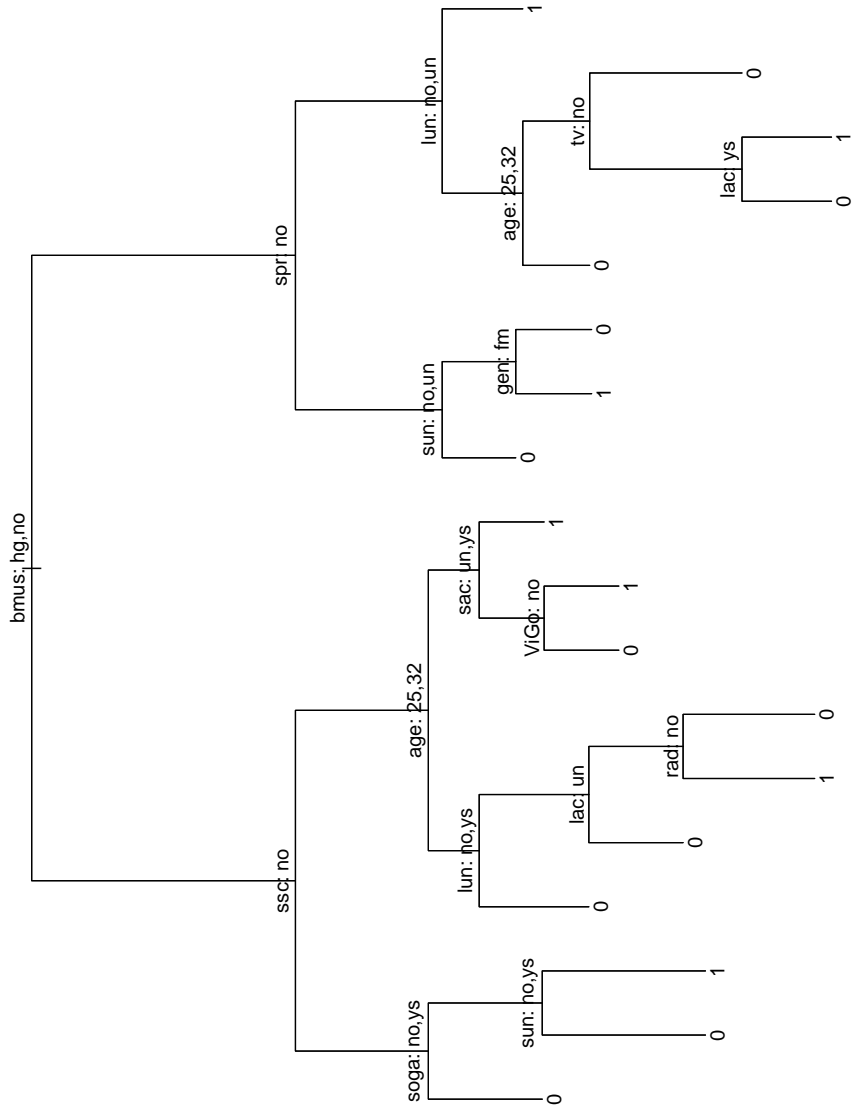


Figure 6.2: Pruned classification tree for the classification between purchase and non-purchase classes of the willingness to pay for the unlimited following slot (ww1)

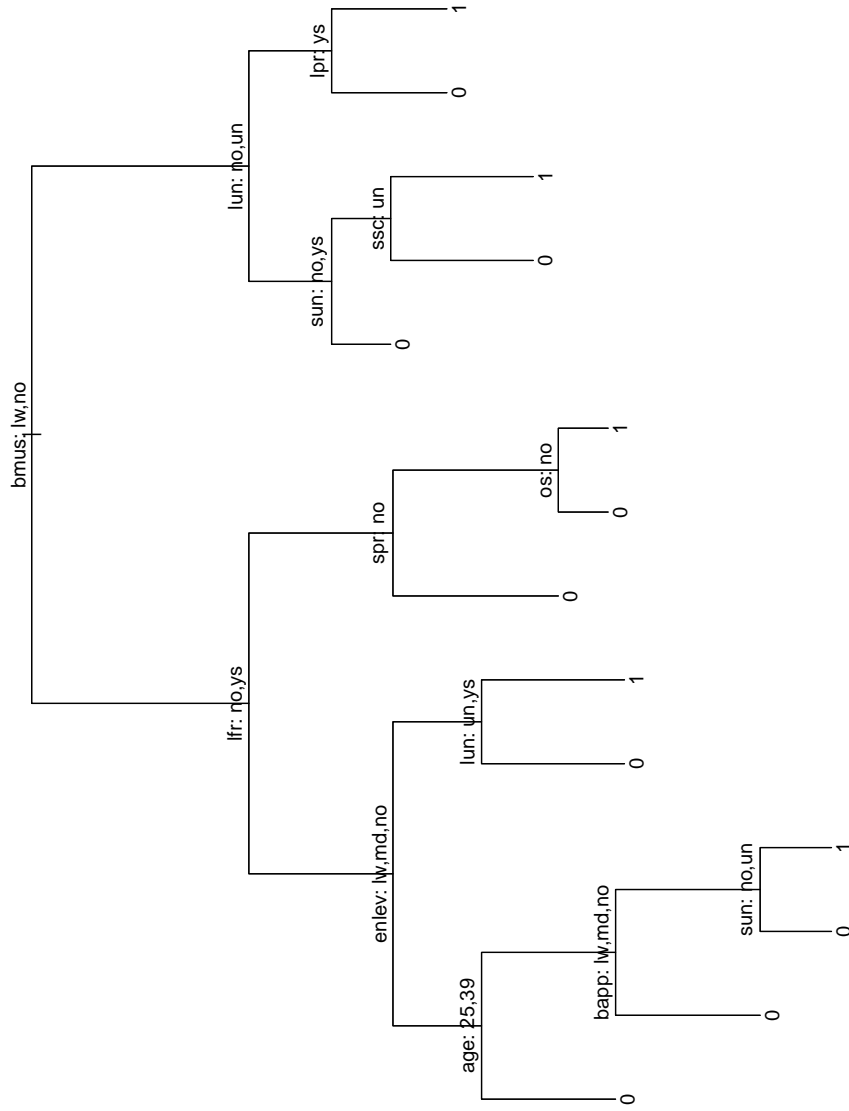


Figure 6.3: Pruned classification tree for the classification between purchase and non-purchase classes of the willingness to pay for the advanced profile (ww2)

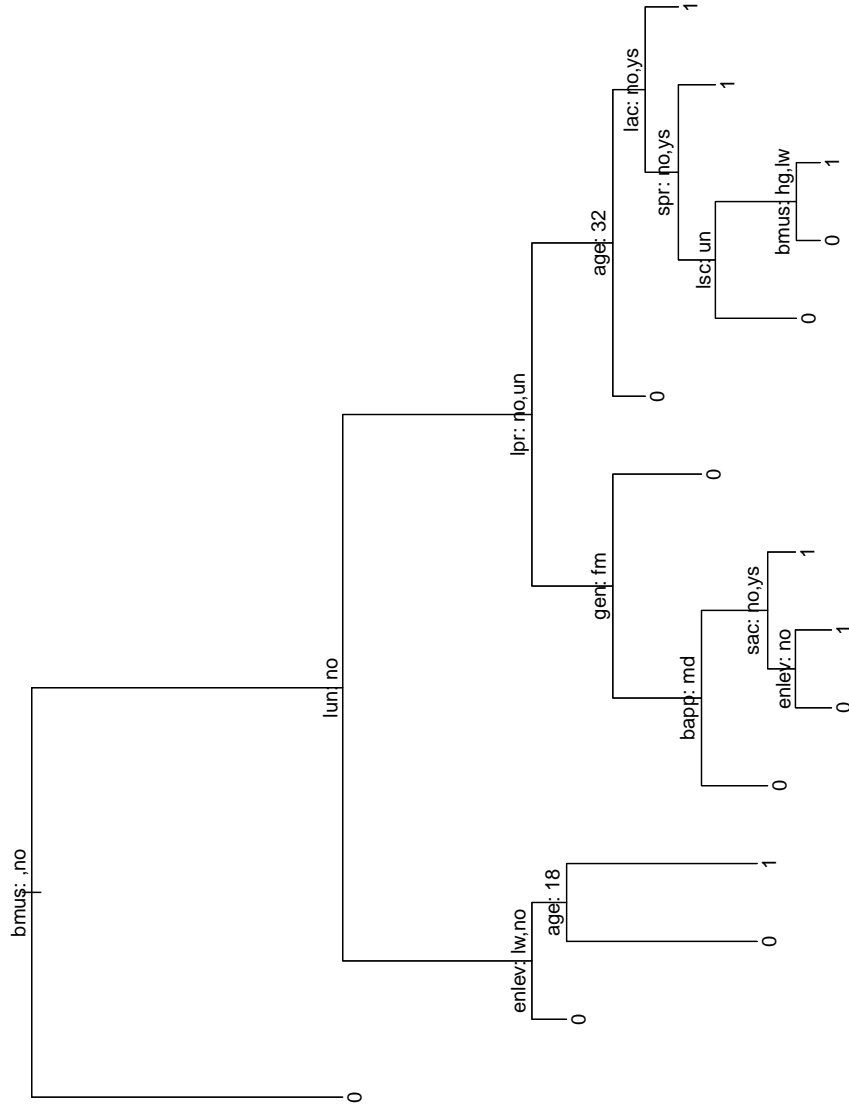


Figure 6.4: Pruned classification tree for the classification between purchase and non-purchase classes of the willingness to pay for the extended range (ww3)

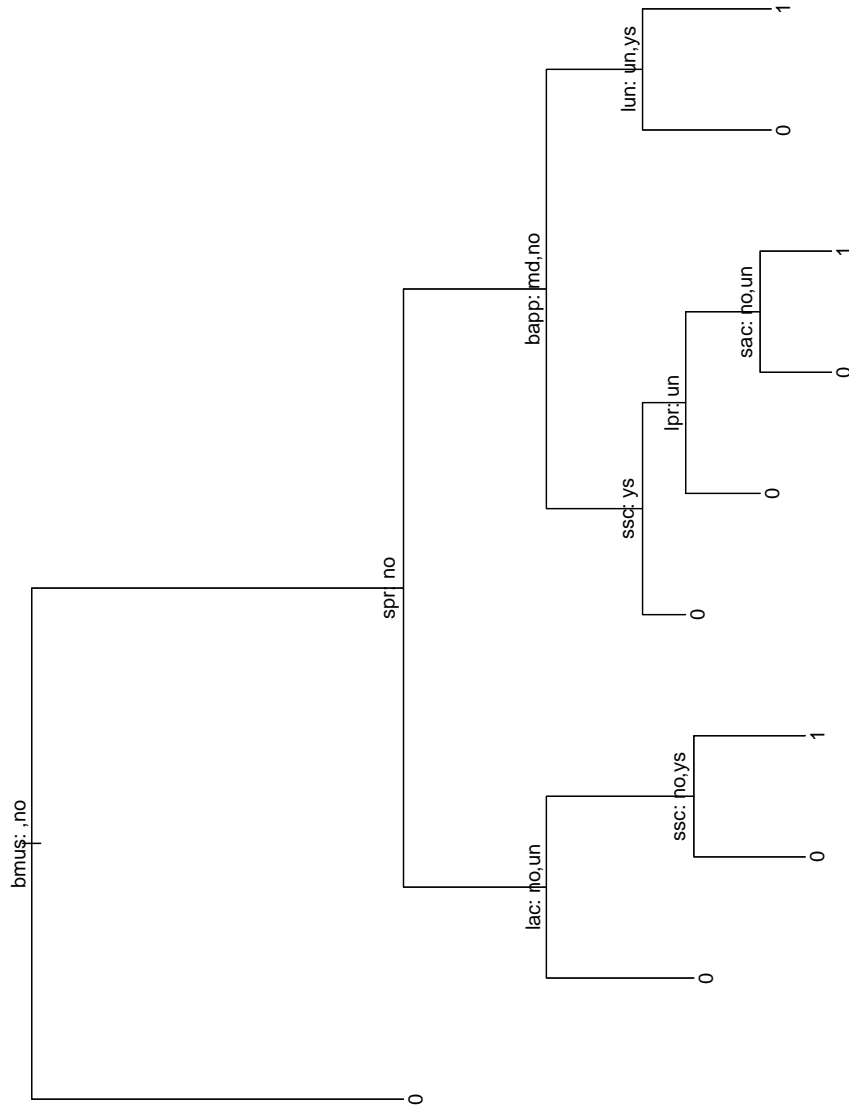


Figure 6.5: Pruned classification tree for the classification between purchase and non-purchase classes of the willingness to pay for the exclusive live music streams (ww4)

7 Model performance assessment metrics

7.1 Confusion matrix

At this stage we aim to provide the model performance assessment for CART and logistic regression models. In our study we deal with binary classification models, since our response variables have only two classes, i.e. purchase and no purchase or a true and a false class. Hence, there are four possible classifications the model can deliver: a true positive, a true negative, a false positive, or a false negative. These scores build up the so called 2×2 contingency table or confusion matrix, which is often used for the model performance assessment, Hamel (2008).

Table 7.1 depicts the elements of the confusion matrix.

		Observed	
		True	False
Predicted	True	True Positive (TP)	False Positive (FP)
	False	False Negative (FN)	True Negative (TN)

Table 7.1: Confusion matrix

The cases that lie on the major diagonal correspond to correct classifications, i.e. true positives and true negatives or in other words true cases which were classified as true and negative cases classified as negative. If the secondary diagonal of the confusion matrix contains values then these signify model errors. False positives or *false alarms* correspond to all cases which are negative but were classified as positive, whereas false negatives or *misses* are the cases of class positive, but were classified as negative.

Applying this logic to our data set, consider our questionnaire which seeks to determine whether a person who possesses certain characteristics is willing to buy a virtual good. A false positive in this case occurs when the person tests as buyer, but actually is not willing to buy. A false negative, on the other hand, occurs when the person tests negative, suggesting he is not interested in buying virtual goods, when he actually does want to buy.

On the basis of the confusion matrix we can derive several model assessment metrics. Accuracy is specified as a proportion of correctly classified classes in the total number

of observations.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7.1)$$

Another performance metric is called *precision* and is the proportion of true positive cases in the total of cases classified as positive.

$$precision = \frac{TP}{TP + FP} \quad (7.2)$$

Whereas a *recall* metric calculates the proportion of correctly classified true classes in the total of observed positive classes.

$$recall = \frac{TP}{TP + FN} \quad (7.3)$$

Each performance metric of a confusion matrix delivers only one scalar. Such types of model assessment measurements were proven to be quite a poor summary of the performance of a model, since derived quality metrics such as precision and accuracy depend on the class distribution in the sample, Provost et al. (1998). The left column of the confusion matrix contains the positive classes whereas the right column combines the negative classes, since accuracy and precision metrics are calculated using the values from both columns, they are sensitive to class skewness.

Table 7.2 demonstrates two confusion matrices (left one with skewed classes). Skewed classes samples occur, when the proportions of observed true and false classes are considerably unbalanced. We employ this example in order to prove that the confusion matrix metrics described above are class distribution dependent. In contrast, Receiver Operating Characteristic (ROC) graphs are insensitive to class skew, because these employ strictly columnar ratios.

		observed				observed	
		True	False			True	False
predicted	True	1250	290	predicted	True	1250	29
	False	750	2100		False	750	210

Table 7.2: Confusion matrix without (left) and with (right) class skewness

Accuracy and precision metrics are computed, which suggest that the classification model based on the left confusion matrix has a considerably higher precision rate, approximately 98% in comparison to 81% of the original model. Also, the accuracy values are higher for the model with skewed classes, which equal to 65% and 76%. This means that a model on the skewed classes data delivers higher perceived quality, whereas the fundamental classifier performance does not change. That is why it is suggested that a ROC graph be used instead of traditional scalar performance indicators. In Figure 7.1

point d illustrates both classification models, showing that ROC methodology is class skewness insensitive. Our data is class skewed, since positive-purchase classes occur approximately a factor 10 times more rarely than negative - non purchase classes. For this reason, ROC appears to be more useful than scalar performance indicators for such data sets.

7.2 Receiver Operating Characteristic Analysis

ROC curves are two-dimensional graphs that depict the performance and performance trade-off of a classification model, Hamel (2008). In order to construct a ROC curve, we need to introduce two other metrics of a confusion matrix. *True positive rate* (TPR) corresponds to *recall* metric and *False positive rate* (FPR) is the proportion of negative cases classified as positive in the total of observed negative classes. As previously mentioned both metrics are strictly columnar, meaning in order to calculate them only the values of the same column are used.

$$FPR = \frac{FP}{TN + FP} \quad (7.4)$$

The metric opposite to FPR is called sensitivity and is calculated as $1 - FPR$, whereas TPR is also called specificity. ROC graphs can be constructed by plotting the TPR against the FPR. Having only scalars from the confusion matrix, we receive the points on the graph. Figure 7.1 depicts the important areas on the ROC graph. Points A , B , C illustrate extreme classifiers. Point C denotes the classifier which produces neither any false positives, nor true positives, this means that all observations are classified as negative. In contrast, point B depicts the classifier which however classifies all true positives correctly but at the same time commits also all false positives. In other words, this model classifies each case as positive.

The perfect classifier is given by the point A , at which specificity as well as sensitivity are equal to 100%, meaning that classification contains neither false positives nor false negatives.

The diagonal line $B - C$ illustrates the random performance. A classification model which lies on this line produces as many true positive responses as it produces false positive responses.

All classifiers mapped to the right of the random performance line commit more false positive instances than true positive instances, for example classifier f .

The region above the random performance line is divided by the orthogonal line through point A into conservative and liberal regions.

The classification model d belongs to the conservative performance region, since it produces quite good true positive rates and low false positive rates. Classifier e is in the

liberal performance region, which is characterised by quite good true positive rates, but also relatively high false positive rates.

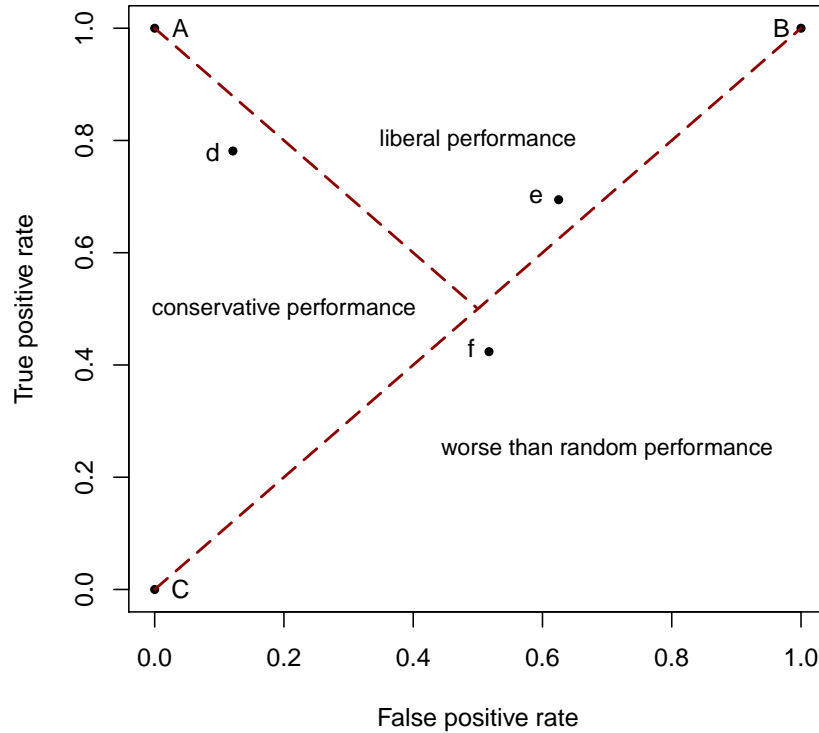


Figure 7.1: Important regions and points of ROC graphs

7.3 Empirical results

Comparing the variables chosen by CART and logistic regression, we can summarise that the CART method tends to sort out more variables for analysis than logistic regression. The overlaps between the two methods can be derived from Table 7.3.

In order to assess and compare the predictive accuracy of the CART and the logistic regression, we divide our data sample into a learning and a test sample, where the test sample consists of 64 observations, which is approximately 10% of the initial sample. The default discrimination threshold of classification models is traditionally set to 0.5, meaning that if the probability is above this cut point, the subject is predicted to be a member of the modelled class. If the probability is below the cut point, the subject is predicted to be a case of the other group.

	Classification Tree	Logistic Regression
ww1	bmus, ssc, sog, sun, age, lsc, lpr, ViGo, tv	bmus, ssc, sog, sun, age, lun, lac, rad, sac, ViGo, spr, gen, tv
ww2	bmus, lfr, enlev, age, bapp, lun, lpr	bmus, lfr, enlev, age, bapp, sun, lun, spr, os, ssc, lpr
ww3	bmus, lun, int, lfr, lsc, sfr, ssc, ViGo	bmus, lun, enlev, age, lpr, gen, bapp, sac, lac, spr, lsc
ww4	bmus, ssc, lpr, sac, gen, int, fr, lsc, sog, enlev	bmus, spr, lac, ssc, bapp, lpr, sac, lun

Table 7.3: Significant variables used in CART and logistic regression analysis, the common variables are marked in blue.

Function `predict.tree` in R produces both a discrete classifier and a vector of probabilities for classes. By verifying the threshold of the probabilistic classifier and computing TFP and FPR of the performance model at each threshold level, we are able to construct the ROC curve. The curve is drawn from left to right, starting with high decision thresholds and ending with lower decision thresholds. For this reason the left side is called conservative and right side is denoted as liberal.

Figures 7.2, 7.3, 7.4, 7.5 demonstrate the ROC curves of CART classification models, in order to provide direct comparison the ROC curves on the basis of logistic regression models are illustrated on the right side. The graphics comprise the predictive ability of classification models both in in-sample and out-of-sample settings.

From the ROC curves of the CART models, we can conclude that all four models in in-sample setting deliver moderate results. The average TPR equals 60% corresponding to FPR of 20%. For all four models the ROC curves lie partially or entirely under the random performance diagonal. In order to compare the performance of different classification models, one can compute the *Area Under the Curve* (AUC) coefficient. AUC is used when a general measure of predictive ability is of interest. The AUC value can range between 0 and 1, because the AUC is a portion of the area of the unit square. The AUC can be calculated by using an average of a number of trapezoidal approximations. One should take into account the random performance diagonal line, which has an area of 0.5, hence, the AUC should be at least greater than 0.5. We calculated the AUC values for CART in an in-sample setting, which are 0.76, 0.77, 0.73 and 0.70. For the CART out-of-sample setting of the ww1 and ww4 the AUC values only slightly surpass the critical value of 0.5, whereas for two other models the ROC

7 Model performance assessment metrics

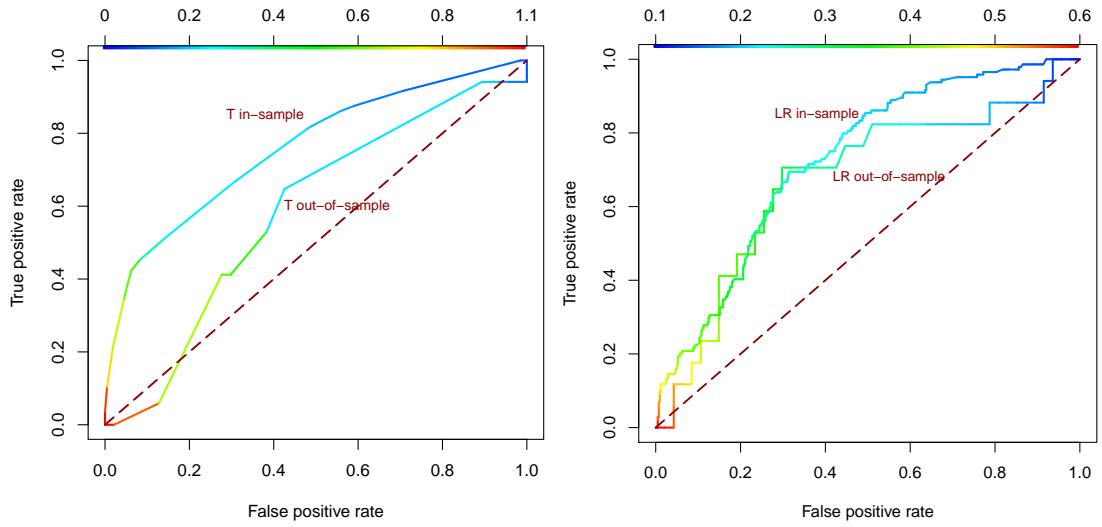


Figure 7.2: ROC for the unlimited following slot (ww1) with CART (left) and LR (right)

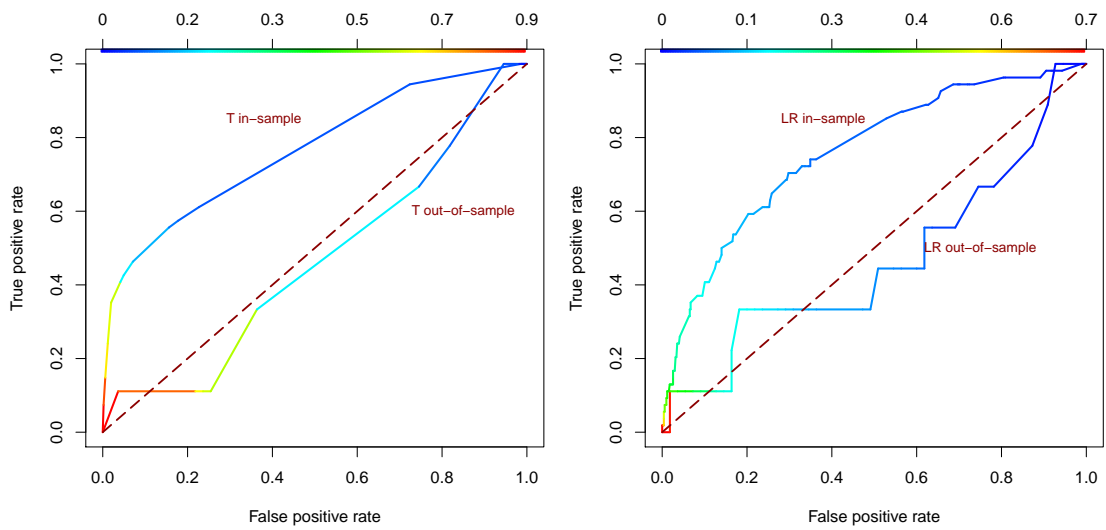


Figure 7.3: ROC for the advanced profile (ww2) with CART (left) and LR (right)

7 Model performance assessment metrics

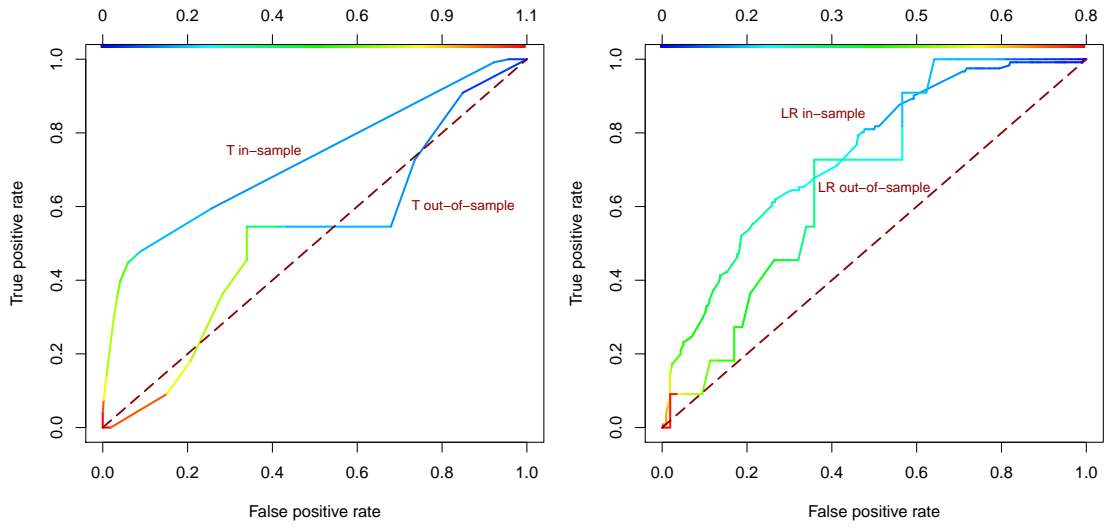


Figure 7.4: ROC for the extended range (ww3) with CART (left) and LR (right)

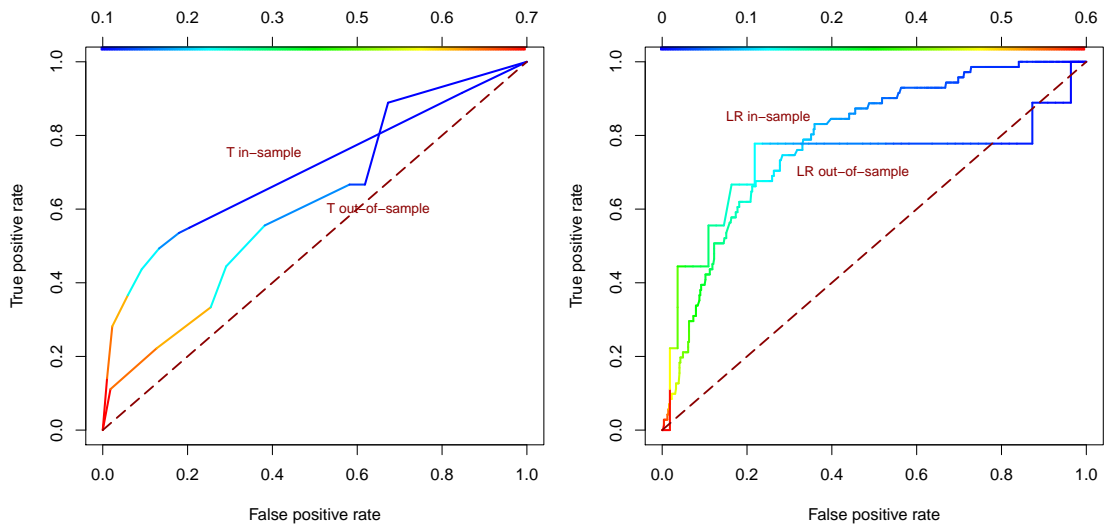


Figure 7.5: ROC for the exclusive live music streams (ww4) with CART (left) and LR (right)

variable	category	learning sample	test sample
gen	female	62.9%	43.8%
age	18-24	55.6%	70.3%
	25-31	32.3%	21.9%
	32-45	12.2%	7.8%
bmus	high	3.6%	7.8%
	middle	11.4%	21.9%
	low	24.2%	10.2%
	no	60.6%	59.4%

Table 7.4: Most significant differences between learning and test samples

response variable	category	learning sample	test sample
ww1	yes	25.7%	26.6%
ww2	yes	9.6%	14.1%
ww3	yes	21.6%	17.2%
ww4	yes	12.7%	14.1%

Table 7.5: Willingness to pay rates in learning and test samples

curve lies considerably beneath the random performance diagonal line.

Whereas the in-sample curves measuring the predictive power of the logistic regression models demonstrate almost identical results as the CART models with AUC values of 0.73, 0.76, 0.74 and 0.79 respectively, the predictive ability of the out-of-sample setting of the logistic regression model are characterised by considerably better results in comparison to the CART models, with the exception of the model for ww2, where the ROC curve lies under the diagonal, the AUC values for three other models are: 0.67, 0.68 and 0.73.

In order to explain why the out-of-sample performance of CART models is rather poor, we compare the test and learning sample descriptive statistics, to check for significant differences, which can be the reason for the results. Table 7.4 illustrates the descriptive statistics of the variables with the most prominent variations for learning and test samples.

Table 7.5 depicts the willingness to pay rates in learning and test samples, which do not considerably differ in hypothetical settings.

Considering the results, there are no significant differences in the two samples, which can be responsible for the unsatisfactory out-of-sample results of the CART models. The exceptions are the variables *gender*, *age* and *monthly budget for music*. Since the first variable is not significant neither in the logistic regression nor in the CART analysis and the distribution of gender is not decisive for the classification model results,

the age structure of the test sample is characterized by the higher percentage of the youngest respondents, but the age variable was significant only for two of the four models. Monthly budget for music is distributed differently between the groups of respondents who spend money on music, although the percentage of those, who do not spend money on music is equal.

Concluding these findings, the poor out-of-sample results cannot be explained due to the heterogeneity of the samples. However, the high misclassification rates of the pruned classification trees, from a minimum of 8% for the second model to the maximum of 19.4% for the first model, the third and fourth trees have misclassification rates of 16.2% and 11% respectively, could be the possible cause of the unsatisfactory out-of-sample predictive ability of the CART models.

8 Conclusion

Our results provided evidence that the direct survey employing CVM of the willingness to pay for virtual goods cannot be used as a unique source for pricing decisions, since the hypothetical responses do not reveal the real purchase patterns. the CVM study pointed out that hypothetical responses highly overstated willingness to pay rates, which are in fact 3 – 4 times higher than the market benchmark.

Considering this fact, the additional usage of the certainty question is further suggested. The hypothetical bias was partially mitigated due to the certainty question, however, since the real market study was out of scope of this work, in order to determine whether the usage of certainty questions are legitimate in the case of virtual goods, further empirical research is demanded and an appropriate survey with real purchase obligations. It was ascertained that the monthly budget for music is the most important variable in all four models in logistic regression as well as in the CART analysis, although, it is possible that other variables exist, which are not considered as covariates in the classification models, but possibly have a significant influence on the WTP decision.

The model performance assessment metrics suggest that the logistic regression to possesses better predictive power than the CART model in an out-of-sample setting. Although, logistic regression also delivers only moderate results.

The absence of the material component of the virtual goods, makes the assessment of WTP even more problematic than for material private goods. More accurate empirical research that would combine hypothetical and real WTP decisions and actual interaction with virtual good are essential to provide more reliable results in determining WTP for virtual goods.

Bibliography

- Becker, G. M., Degroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9:226–232.
- Block, B. (2011). Smartphones gewinnen an fahrtwind in deutschland. Technical report, comScore, Inc.
- Blomquist, G. C., Blumenschein, K., and Johannesson, M. (2009). Eliciting willingness to pay without bias using follow-up certainty statements: Comparisons between probably/definitely and a 10-point certainty scale. *Environmental and Resource Economics*, 43(4):473–502.
- Blumenschein, K., Blomquist, G. C., Johannesson, M., Horn, N., and Freeman, P. (2008). Eliciting willingness to pay without bias: evidence from a field experiment. *The Economic Journal*, 118:114–137.
- Blumenschein, K., Johannesson, M., Blomquist, G., Liljas, B., and OConor, R. (1998). Experimental results on expressed certainty and hypothetical bias in contingent valuation. *Southern Economic Journal*, 65(1):169–177.
- Breidert, C., Hahsler, M., and Reutterer, T. (2006). A review of method for measuring willingness-to-pay. *Preprint to appear in Innovative Marketing*.
- Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks.
- Castronova, E. (2002). On virtual economies. Technical report, CeSifo Working Paper No. 752.
- Champ, P. A., Bishop, R., Brown, T., and McCollum, D. (1997). Using donation mechanisms to value nonuse benefits from public goods. *Journal of Environmental Economics and Management*, 33:151–62.
- Cummings, R. G. (1997). Are hypothetical referenda incentive compatible? *Journal of Political Economy*, 105:609–621.

Bibliography

- Cummings, R. G. and Taylor, L. O. (1999). Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *The American Economic Review*, 89(3):649–665.
- Denegri-Knott, J. and Molesworth, M. (2010). Concepts and practices of digital virtual consumption. *Consumption Markets & Culture*, 13:2:109–132.
- Hamel, L. (2008). *The Encyclopedia of Data Warehousing and Mining*, chapter Model Assessment with ROC Curves. Idea Group Publishers.
- Harrison, G. W. and Rutström, E. E. (2008). *Handbook of experimental economics results*, chapter Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods, pages 752–766. Elsevier.
- Hosmer, D. W. and Lemeshow, S. (1989). *Applied logistic regression (Wiley Series in probability and statistics)*. Wiley-Interscience Publication.
- IFPI (2011). Digital music report 2011. music at the touch of a button. Technical report, International Federation of the Phonographic Industry.
- Johannesson, M., Blomquist, G. C., Blumenschein, K., Johansson, P.-O., Liljas, B., and O’Connor, R. M. (1999). Calibrating hypothetical willingness to pay responses, journal of risk and uncertainty. *Journal of Risk and Uncertainty*, 8:21–32.
- Johannesson, M., Liljas, B., and Johansson, P.-O. (1998). An experimental comparison of dichotomous choice contingent valuation questions and real purchase decisions. *Applied Economics*, 30:643–47.
- Lehdonvirta, V. (2008). Virtual worlds dont exist. In *Breaking the Magic Circle*.
- Lehdonvirta, V., Wilska, T.-A., and Johnson, M. (2009). Virtual consumerism: case habbo hotel. *Information, Communication & Society*, 12:10591079.
- List, J. and Gallet, C. A. (2001). What experimental protocol influence disparities between actual and hypothetical stated values? *Environmental & Resource Economics*, 20:241–254.
- List, J. and Lucking-Reiley, D. (2000). Demand reduction in multiunit auctions: Evidence from a sportscards field experiment. *American Economic Review*, 90(4):961–972.
- Little, J. and Berrens, R. (2004). Explaining disparities between actual and hypothetical stated values: Further investigation using meta-analysis. *Economics Bulletin*, 3:1–13.

Bibliography

- Loomis, J., Brown, T., Lucero, B., and Peterson, G. (1996). Improving validity experiments of contingent valuation methods: Results of efforts to reduce the disparity of hypothetical and actual willingness to pay. *Land Economics*, 72(4):450–461.
- Magid (2010). Magid report 2010: Market for mobile virtual goods. Technical report.
- Mitchell, R. C. and Carson, R. T. (1989). *Using surveys to value public goods: the contingent valuation method*. Resources for the Future.
- Nielsen (2010). Music mobile apps and music streaming services. identifying the consumers and tapping into the new opportunities. Technical report, Nielsen.
- Provost, F., Fawcett, T., and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 445–453.
- Reuters (2011). Zynga draws fewer paid players than expected. *Reuters*.
- Schneider, A. (2008). Virtual item monetization: A powerful revenue opportunity for online game publishers and virtual world operators. *Live Gamer*.
- Sheth, J. N., Newman, B. I., and L., G. B. (1991). Why we buy what we buy: A theory of consumption values. *Journal of Business Research*, 22:159–170.
- Skiera, B. and Revenstorff, I. (1999). Auktionen als instrument zur erhebung von zahlungsbereitschaften. *Zeitschrift für betriebswirtschaftliche Forschung (ZfbF)*, 51:224–242.
- Stelzer, D. (2004). *Entwicklungen im Produktionsmanagement*, chapter Produktion digitaler Güter, pages 233–250. Hans Corsten.
- Timofeev, R. (2010). *Statistical Aspects of Stock Picking and Risk-Averse Behaviour*. PhD thesis, Humboldt-Universität Berlin.
- Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance*, 16(1):8–37.

Appendix

ONLINE SURVEY

Willingness to pay for virtual goods

Dear friends,

The essential part of my master thesis is a survey about the willingness to pay for virtual goods on the example of an upcoming social music service for smartphones. I would like to seek your support and kindly ask you to answer a short questionnaire.

The survey takes less than 10 minutes to complete. I thank very much in advance everyone who takes part!

Please, consider there are no right and wrong answers and only your personal opinion and experience do matter.

Among all participants 2 urbanears headphones in colour of your choice urbanears.com of 40 value each will be lottery drawn.

Faithfully yours,

Polina Marchenko

A note on privacy This survey is anonymous. The record kept of your survey responses does not contain any identifying information about you unless a specific question in the survey has asked for this. If you have responded to a survey that used an identifying token to allow you to access the survey, you can rest assured that the identifying token is not kept with your responses. It is managed in a separate database, and will only be updated to indicate that you have (or haven't) completed this survey. There is no way of matching identification tokens with survey responses in this survey.

Demographic

Please fill in some basic information about your person.

1. Gender
 - Female
 - Male
2. Age
 - less than 18
 - 18 - 24
 - 25 - 31
 - 32 - 38
 - 39 - 45
 - 45 +

Smartphone usage

Please fill in some information about your smartphone usage patterns.

3. Do you have iPhone?
 - yes
 - no
4. What is your monthly budget for mobile applications?
 - 0 €
 - less than 5 €
 - 5 - 10 €
 - more than 10 €

Music affinity

Please fill in some information about your music affinity.

5. How do you explore new music?
 - Internet

Social music service - wahwah.fm

wahwah.fm - is an upcoming service for smartphones enables you to enjoy the live music sharing experience and more.

How does it work?

You can decide either being a Listener of other broadcasts or an active music Broadcaster yourself.

Listener and Broadcaster listen exactly the same music at the same time.

- as Listener, you can tune in music stream of any Broadcaster in your neighbourhood.
- as Follower you can tune in music stream of your favourite Broadcasters also when they are not nearby any more.
- as Broadcaster you can create your own music station and easily share your music, get fans, receive feedback.



Screenshots of the iPhone application

Appendix

9. Basic version - you can follow 5 broadcasters for free. Monthly subscription to the unlimited number of broadcasters is possible. Would you buy the monthly subscription to the unlimited number of broadcasters for 0.79 €?
 - yes
 - no
10. How sure are you about buying (not buying) the monthly subscription to the unlimited number of broadcasters?
 - probably sure
 - definitely sure
11. wahwah.fm is a location linked music search engine. As Broadcaster you compete for attention with other broadcasters on location. You can achieve more visibility on the map and in the list with advanced profile. Would you buy a monthly subscription to the advanced profile for 0.79 €?
 - yes
 - no
12. How sure are you about buying (not buying) the advanced profile?
 - probably sure
 - definitely sure
13. Basic version - you can enjoy broadcasts in the city you are located in. With extended range listenership you get access to thousands of broadcasters in other German cities. Would you buy a monthly subscription to extended range listenership for 0.79 €?
 - yes
 - no
14. How sure are you about buying (not buying) the extended range listenership?
 - probably sure
 - definitely sure
15. Virtual ticket gives you 24h exclusive access to the live music stream. You can attend multiple closed music events in one day. Would you buy a virtual ticket for 2.99 €?
 - yes
 - no

16. How sure are you about buying (not buying) the virtual ticket?

- probably sure
- definitely sure

Virtual Games Experience

17. Have you ever participated in social games, for example Farmville, Mafia Wars on Facebook & Co.?

- no
- uncertain
- yes

18. How would you assess your engagement level in social games?

- no engagement
- low (I play very rare)
- middle (I play sometimes when I check into my profile)
- high (I play each time when I check into my profile)

19. Have you ever spent money on virtual goods, such as avatar's accessories, virtual animals, gifts etc.?

- yes
- no

Thank you for your participation!

In order to participate in the lottery, please send email to: polina.marchenko@yahoo.com. If you have iPhone and you want to become one of the first to enjoy the advantages of wahwah.fm, you can apply for the private Beta testing. Go to wahwah.fm/beta.