UNIVERSITY
OF TWENTE.

FORECASTING THE TRANSACTION FEES PER CLIENT

In search for client segmentation

Abstract

How to predict the fee per year of a new investors and give this investor insides in its expected transaction behaviour, given some characteristics. A research conducted within company X on the execution only clients.

Groot, R.W. (Rogier)

Preface

Herewith we present you the master thesis, conducted for the master Industrial Engineering and Management, specialization Finance. The research was conducted within company ‘X’ at the department ‘Y’ in the team of ‘Z’. We hereby would like to thank department Y and the team ‘Z’ for giving me the opportunity to do the master thesis on their business unit.

We would like to thank in special the mentor within company X, Wouter Purmer, for all his stimulations and the nice talks to improve the research. We would like to thank Dr. Berend Roorda and Drs.ir. Toon de Bakker for their feedback on the thesis during the process of writing and for being critical and stimulating when needed. We further would like to thank Roy Eijkelboom for all his help with gathering the right data and for checking the queries and scripts. Last but not least we would like to thank the team of Z for the very interesting six months in their team and for the daily stimulation within company X.

After six very interesting months on the business unit within department Y and the months afterwards to finish the thesis, we hereby present you the results of the research.

Index

Management Summary

This master thesis has the following problem statement "*How could X increase its user friendliness with regards to the information on costs of investments for investors, keeping in mind regulatory needs and how should X show these costs omnichannel?"* After some brainstorming within X the idea was trying to build a predictive forecasting model for those new investors interested in execution only. The idea is that new investors get insights in the cost for transactions as well as a prediction for the service fee. The overall was thus to forecast the transaction fees by client segmentation.

After reviewing literature, some relations were found for segmentation, for example it suggested that male investors tend to invest more than female investors do, thus resulting in more cost. Do men have more trades per year, compared with women. It suggests that older people trade less than younger ones do. Investors trade more often when their level of wealth increases. The more experienced traders are, the more they tend to trade. People with lower education seem to invest more in stocks and people with higher education levels trade more than lower education levels. This all suggests that one should be able to distinct the cost investors make per year based on these characteristics.

After three steps the dataset was collected within X. The first dataset we started with, was that with generic data and some of X segment classes. Above that we had the investment amount and the summed data of transactions per category. With the second dataset, it was possible to connect the account number of the clients with the knowledge and experience tests of these investors, to retrieve their education level and experience with investing. The last dataset was collected with all the transaction data. After summarizing these per account number all three variables were connected with each other.

The next step was to deal with the outliers of the dataset and with the use of the Winsor method, the outliers were dealt with. Via the use of a decision tree we then decided which factors seem to predict and distinct the cost per year the most. After some tests four predictive factors remained to investigate. Wealth, gender, general experience with investing and education remained as the four factors for the model. After testing the different combinations, the model with wealth, gender and education was found to be the best model with nearly 28 percent predictive power. The predictive power was calculated with the use of the risk estimate and deals with the explained variance of the model. Second best was the combination of the four factors, with 27,5 percent explained variance. Concluding that 28 percent predictive power is somewhat low, we still wanted to take a look what the performance of brokers in general. We have modelled the most known brokers against each other.

We would suggest X to continue the investigation on these factors in the future. The decision tree gives insights in which persons X could target as potential new investors and how it should place itself in the market. The model also keeps in mind the categories in which is invested, the stock exchanges used and the exchange rates. This could also give insight in what investors in the sub selections tend to find most interesting compared with the other selection and groups. Although the predictive power of the model isn't as high as desired, the dataset and SPSS output provided some insights on the literature used. In the dataset provided by X, based on the transaction cost paid per year, significant differences (95 percent confidence interval) where found between gender, education level and wealth.

Male investors tend to pay more transaction cost per year than female investors do. The cost per year paid, based on the education level is that investor with a lower education level tend to pay more than investors with a higher education level. The distinction made is here between primary, secondary and university (or higher) education. Also higher levels of wealth are significant different than the ones with less wealth. This suggests, as the literature also stated, that increases in wealth also increases the number or at least the value of the transactions done.

1 Introduction and approach

In the following chapter we will introduce the problem that arose within X and will introduce the project. After the problem statement the research question is presented.

1.1 Introduction

With the current interest rates below 1% at every banking facility in the Netherlands and the Eurozone, clients of banks tend to search for alternatives. One of these alternatives could be investing in the stock market. In 2014 a 20% increase was noticed in the number of people investing on the stock exchange, compared to 2013 (Millward Brown, 2014). These new clients started investing in the stock exchange, since they had resources available and the interest rate was too low in their opinion. Their willingness to take more risk caused them to invest in the stock market. Also the media reported that in 2015 there was an increase of 50% of retail investors (Rezelman, 2015).

Although X already offers a platform where clients can invest in the stock exchange, it wasn't their main focus. But with the current stimulus program of the European Central Bank, with interest rates around zero percent, banks are offering their clients alternatives. Although there exists a lot of competition in the brokerage world, X also decided to work on the improvement of their stock market facilities.

1.1.1 Current project

One of the selected segments to grow in, are the brokerage activities. Investing provides their clients an alternative in the current low interest rates environment. The (investment) webpages on X.nl, including those of brokerage, have got several upgrades and the goal is to make investing as easy as possible for the clients. The Y department is working hard on the transition to make investing omnichannel, for as well the retail as the business clients. Omnichannel is a multichannel approach that seeks to provide the customers a seamless experience whether the customer is online on a desktop or mobile device, or for example by 'regular' telephone. One of the projects is the development of an investment application, for mobile as well as for tablets. This application needs to satisfy some regulatory needs which will be worked out in 1.1.3.

1.1.2 Idea and flow of the application

The (regulatory) needs within the app are next to the general agreement and acceptation of the general terms, testing the client's knowledge of investing and inform him about the tariffs. These costs need to be transparent, simple, clear and have to be conform the regulations of the AFM.

The application starts with several questions about the investor and his tax registration. Further in the application the intention of the investor is measured. The idea is that after the investor has chosen its product, the costs are shown. These costs need to be conform the regulations stated by the AFM.

1.1.3 Problem with cost structure X

With the current regulation of the AFM all the investment brokers, including banks, need to be transparent, simple and clear. This results in a big list of all kind of costs which are shown in Figures 1,2 and 3. In Figure 1 the basic fee is shown. Figure 2 provides the variable service; Service fee is collected every quarter based on the average invested amount at the end of each month.

Figure 3 shows the cost of different products. Investment funds and trackers are free of charge at X. Stocks, obligations, sprinters and structured products have a fixed fee of €4 plus 0,04% of the invested capital (with a maximum of €150). Options are priced €2,25 per contract. If the transaction is in a foreign country, the client occasionally might need to pay a 'Stamp Duty' or 'transaction tax'. Foreign obligations have a minimum transaction cost of €50. With orders in another currency, X calculates 0,25% of the middle bid-ask spread of that currency.

Figure 4 shows the cost of stock-, obligation-, index-, tracker-, real-estate-, liquidity- and alternative investment funds. All these funds have their own percentage which they deduct over the invested capital. X doesn't make profit over these funds, although these cost need to be shown to the investors.

All these cost are directly related to every investment account when one is investing in one of these products. Next to the tariffs page X.nl shows more costs that relate to the investment account somehow. Think of interest rate over the cash, costs involving forced sale of options/stocks or the cost of having debt on the account.

The overview of the webpage is found to be not very clear and simple as an investor might hope it would be. Although X is transparent about the cost of investing, all the exceptions make it quite unclear and the webpage doesn't give a nice overview. But the real problem occurs when one should mention all these cost in an opening flow on a mobile phone. On the webpage, clients already get lost in all the different costs per product and therefor miss the clear overview in costs that is needed. So think about how new clients would feel on their mobile phone?

Costs are important for X's clients as they wish to see their 'real' investment return (net profit). They should be able to know their cost when investing. But not all the investors invest in the same products, they differ for example on: The investment categories chosen (i.e. stocks, options etc.), the amount of cash invested, the number of transactions per year, the cost of indirect products (i.e. investment funds) versus the direct cost (i.e. basic fee).

1.1.5 Problem statement and purpose thesis

As described in 1.1.4 the amount of information required is huge and the transparency over the cost of investing at X isn't as clear as it should be. As X is developing an application to make investing omnichannel, the huge loads of information and the unclearness with this is far from the desired situation. X wishes there is a better way to show all these costs in the application than just simply stating the facts. The results from this research to the costs for the application might also be applied to the general webpage in order to simplify the page as well. This research will first categorize the investors and then will investigate if there is a relation between these investors and the cost structure associated with their execution only profiles. The problem statement that will be worked out is:

How could X increase its user friendliness with regards to the information on costs of investments for investors, keeping in mind regulatory needs.

2 Research approach

This section describes the research approach of this study. The central problem statement will be worked out, as well as the sub research questions. The general outline of this research will follow after the research (sub)questions are determined.

2.1 Central Research Question

From the description of the subject and the problem that arose during the purpose of this study, mentioned in the previous paragraph, the following problem statement is established.

How could X increase its user friendliness with regards to the information on costs of investments for investors, keeping in mind regulatory needs?

To answer the problem statement several sub research questions have been established. Via the research structure mentioned in section 2.2 it will briefly be explained how every chapter is built up and what is needed to answer the sub question.

2.2 Sub research questions and Methodology per research question

In addition to the above mentioned problem statement, five sub research questions are formulated to support the main research question and helps to answer the main research question.

Research question 1:
Is it possible to characterize the different types of investors based on literature research?

	Methodology	Research	Required data	Notes
1.	Characterize the investor clients and gain insides on their investment strategy			
a.	Literature research of influencing factors for investment in securities	Literature research	Literature	
b.	Determine which literature is within scope	Internal research & discussions X	Literature	Out of scope versus in scope
c.	Decide which literature is worked out	Literature research		

This study strives to improve the information on the cost structure and the clearness over the costs of investment for the clients that invest via X. To categorize these investors, first some literature is reviewed for some guidance on the characterization. If it is possible to categorize the investors into some groups, their investment strategy will be reviewed and some interesting facts will be stated. To do so a good overview of the customers of X is needed. So to answer this question, we will do a literature research, discuss some interesting insights and after discussing what is relevant place some subjects out of scope or destinate them for further research.

Research question 2:
Which factors, influence the trading behaviour of clients, can be found within X?

	Methodology	Research	Required data	Notes
2.	Investigate data, that can be found, which influence the trading behaviour of clients?			
a.	Dataset scan and collection	Data research	Data set clients "Product"	
b.	Accordance between data and literature	Data research	Data set inspections of clients with "Product"	Combine literature and data
c.	Make data applicable for model and comparison	Data	Literature/ internal	Make all data available in one program

To investigate and determine which factors really influence trading behaviour of investors at X, we need to check which outcomes of the literature study are also applicable to the dataset of X. First we will review the literature with the data research, as some will not be available within X. We also expect not to get all the needed data into one sheet. So after collecting all needed data it should be adjusted to make it applicable for comparison. After this step the data is applicable in one program to further analyse it.

Research question 3:
Is it possible to make an estimation of the cost a client has had, based on multiple factors and are these cost useful to make a yearly prediction of the expected costs?

	Methodology	Research	Required data	Notes
3.	Gain insights in the historical cost a client has had and determine if these historical cost are useful to make a reliable model of future cost?			
a.	Select relevant regression method	Literature and statistics books	Literature/Courses	Information found on Blackboard.
b.	Research of each variable individual	Data Research		Determine via SPSS the influence of dimensions.
c.	Insights in historical cost	Data research	Internal "Product" data	Construct table overview on most influencing dimensions
d.	Develop future cost scenarios per client based on proven variables	Data research/ Literature research	Client trade data / literature	Different scenarios for estimation. (all in cost, on sub categories)
e.	Review constructed scenarios	Data research	Literature/ internal	Review scenarios via simulations or in SPSS

The influence of costs on the net return will be based on the different factors the investors have. These costs will be based on historical data of clients within X. We try to determine the effect of costs on their average investment proportion. First we need to determine the influence of all the dimensions and then draw conclusions. Afterwards we determine the historical cost of the different investor groups, based on their characteristics. We investigate if any prediction model is applicable for these investor groups in order to estimate the costs they will have in a year. We need to test the model on its reliability and these tests are conducted with the use of SPSS. We also investigate if it is possible to run it in a simulation to calculate different scenarios for the costs.

Research question 4:
How does X perform, based on a simplified method, in comparison with its competitors?

	Methodology	Research	Required data	Notes
4.	How to show the cost of investing in securities and when is cost the dominating factor in selecting a broker?			
a.	Develop competitor broker overview	Webpage information	Public data of cost	Show cost for transactions

First we need to make sure what all the competitors charge on different products and exchanges. So first we will compose a competition overview.

2.3 Research Methodology per research question

In Figure 5 the conceptual framework of the research is stated. The top diagram gives the relation between investing, return and cost. The return on investments, investors have, are influenced by the cost one pays. Under the dotted line, the process of developing the prediction tool is visually shown.

It starts with the investigation of the available academic literature. Lots of studies have been performed on performance and characteristics. The relation between return and cost is clear and in the process of getting the relevant literature research with return and cost will be applicable.

After gathering the (relevant) literature, a literature matrix is developed to distinct which paper addresses which subject(s). After working out the different influencing variables of trading and cost, a decision is made, based on common sense and on usefulness. Those that were considered useful are worked out and those that are interesting, but out of scope are placed in the appendix.

After determining the factors, the next phase is to test whether they also influence the clients of X with ‘product’. The data available will be used to test the most influencing factors. Although some influencing factors are difficult to determine via the data. But in accordance with XX (Questionnaire partner of X) and X, the conclusion came that the raw dataset was that pure, a questionnaire would make the research and outcomes less trustworthy. So only the applicable factors found, that are also found in the data within X, are used.

After gathering the data of clients with ‘Product’, the data needs to be prepared for comparison as we will conduct the review over the period of 2014, 2015 and 2016. The choice of 2014 is because of the tariff changes that became active on the first of January 2014. For the comparison we will only use the private clients and will exclude the business customers.

After the data analysis the factors will be tested and will be reviewed whether they also influence the dataset with X’s clients, or that we need to neglect some. For the use of the tool and the process(es) it will be in, it is decided to take not too many variables into account in the prediction. Although the confidence and strength of the predicting value is brought down by this, the simplicity also needs to be taken into account. The final tool needs to be built up by three, four or five variables that result in good confidence. If the fifth variable only slightly increases the confidence and reliability, it will be left out due to simplicity.

The simplicity is of high importance, as X wants to provide this tool as an extra service. The tool should provide in just three or four steps a good estimation of the clients cost. The expectation should not scare the customers away, but help them in the process of deciding that X is the right broker for them. By keeping the model simple, the comparability and use could be connected with competing brokers, to give a prediction of the cost to expect at competing brokers.

After the selection of the most influencing factors and thus the best predicting factors when it comes to cost, these factors will be included in the prediction of expected yearly cost. These factors will characterize the (new) client, and with this characterization, an expected value of cost will follow. The expected cost will be based on the dataset (with 100.000 different clients) that are grouped on these factors.

After reviewing the model and adjusting it where needed, the model will be ready and can be included within the Onboarding flow of opening an account on “Product”, or can be included on the webpage, as a simple tool to estimate the expected cost for a client.

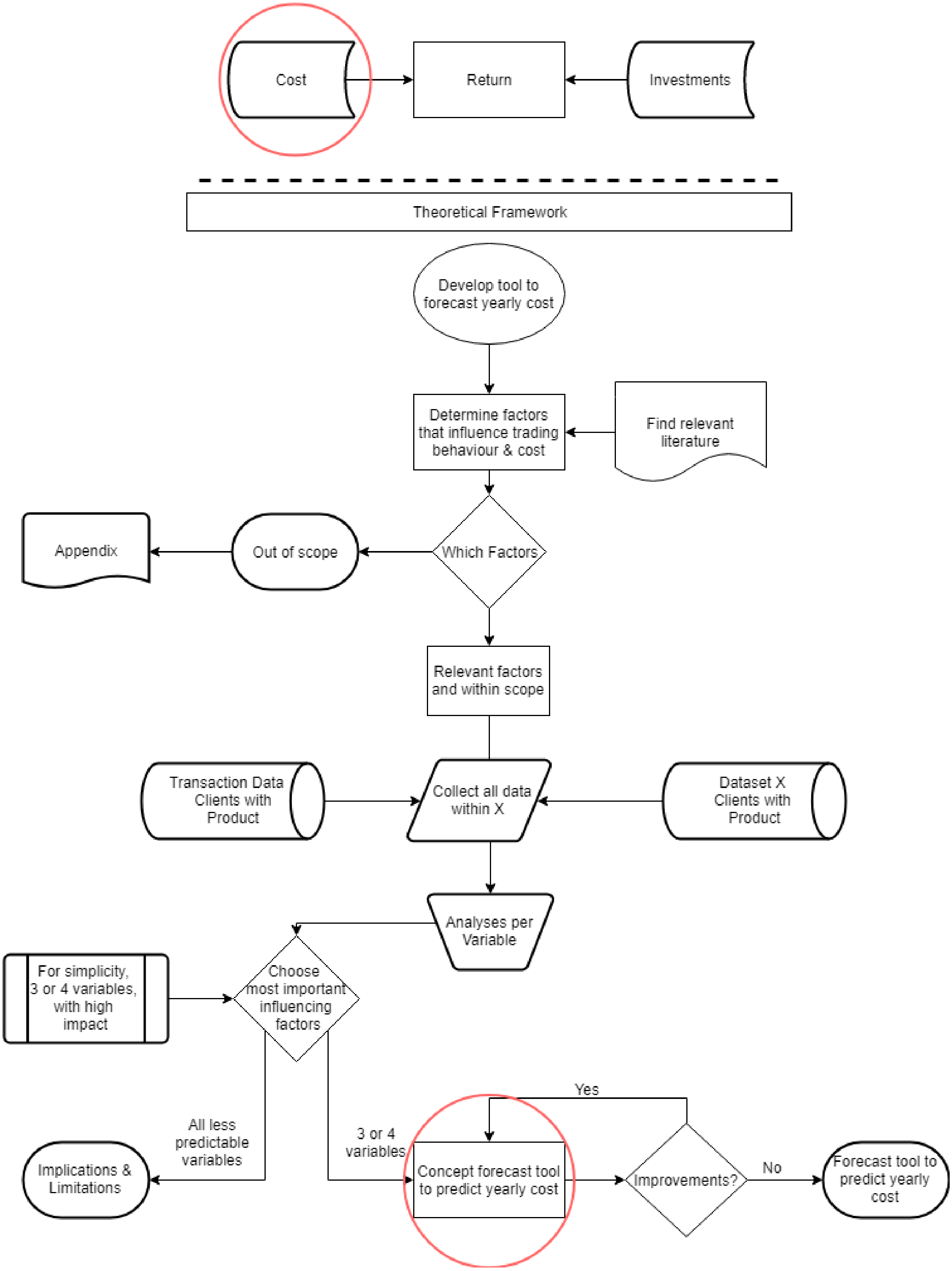
Cost
Return
Investments
Theoretical Framework
Develop tool to forecast yearly cost
Determine factors that influence trading behaviour & cost
Find relevant literature
Appendix
Out of scope
Which Factors
Relevant factors and within scope
Transaction Data Clients with Product
Collect all data within X
Dataset X Clients with Product
Analyses per Variable
For simplicity, 3 or 4 variables, with high impact
Choose most important influencing factors
Yes
All less predictable variables
3 or 4 variables
Implications & Limitations
Concept forecast tool to predict yearly cost
Improvements?
No
Forecast tool to predict yearly cost

Figure 1 - Conceptual Framework

3 Characterize the different types of investors and their strategy

In this chapter the first sub research question will be researched and worked out:

Is it possible to characterize the different types of investors via literature research?

At the end of this chapter we hope to be able to characterize the investors within X and be able to place them in groups. Hopefully also some insights about their investment strategy are found. To reach this, we will work out the available literature first. Afterwards we will develop and restate all hypotheses found in the literature.

3.1 Literature guidance with regards of types of investors

In this paragraph the available literature will be worked out. Although a lot of literature is available, we focus on the literature that discriminates the different types of investors. As Andersen (2013) states: "High-stake investors are, on average, overconfident in their abilities to invest successfully, and they trade more. They have less wealth, are younger, more likely to be men, and have a lower level of education when compared with those with less concentrated portfolios." This already suggests a lot of parameters influence the prediction of costs. In this paragraph the different parameters will be worked out. In which paper, which subject is worked out, can be found in the literature matrix, available in appendix A.

3.1.1 Gender

Anderson (2013) investigates the trading behaviour and the diversification of portfolios, with regards to a lot of influencing variables. When it comes to male versus female investors, high-stake investors are more likely to be male, which is in line with the research of Barber and Odean (2001), who find that men trade more than women. In which stake size is concerned as : "The portfolio value divided by the total risky financial wealth." Which concerns the amount of wealth in under-diversified portfolios. Anderson (2013) finds in its data set that women respond positively to past trading returns and have, on average, a lower sensitivity to stake size. Also women are quite insensitive to losses on paper in the dataset.

Barber and Odean (2007), find, with a subset of the Taiwan stock exchange, again a difference between men and women. Where both men and women prefer to sell winners rather than losers. Men tend to sell the losers at a higher rate, which is in line with Barber and Odean (2001), they find that man trade more than women (although with a subset from a large US discount broker). Barber and Odean (2007) found also that men are somewhat more likely to sell short than women. Grinblatt and Keloharju (2001) also found evidence for the fact that men and women have similar propensities to sell, with a Finish dataset. The greater propensity for men to buy rather than sell would be consistent with men trading more than women. Although Grinblatt and Keloharju (2001) warns for the fact that any gender is consistently a net buyer of stocks relative to the other gender.

Barber and Odean (2001) claim both men and women are overconfident when it comes to trading, and that overconfident investors trade too much. Although men are more overconfident than women and thus men will trade more and perform worse than women. The results found are strongest between single men and single women. By trading much often, men incur higher transaction costs, and consequently earn lower returns.

Dorn and Huberman (2005) found for a subset in Germany in line with Barber and Odean (2001) that younger and male investors trade more aggressively than older and female investors, and also found that older or more experienced and better educated investors hold
less concentrated portfolios (Goetzmann and Kumar, 2002). Also Dorn and Huberman (2005) found that male investors tend to report to be less risk-averse, although not as robust as overconfidence.

Dorn and Huberman (2009) found also that male investors and wealthier investors appear to enjoy dealing with investments more than their female and less wealthy counterparts.

Feng and Seasholes (2007) found for the emerging market of China that the degree of home bias, the fact both men and women over-weight local stocks, is equal among gender. Both sexes tend to invest more in the local stocks. The performances of males versus females is not significantly different. The last result of their emerging market analysis is that men tend to trade more intensively than women before controlling for factors such as number of trading rights. Men hold larger portfolios and make slightly larger trades.

Graham et al. (2009) also found that male investors, and investors with larger portfolios or more education, are more likely to perceive themselves more competent than female investors, and investors with smaller portfolios or less education. They found the relation between overconfident investors tend to perceive themselves to be more competent, and thus are more willing to act on their beliefs, leading to higher trading frequency.

In the clusters of Keller and Siegrist (2006), again men are overrepresented in the risk-seekers cluster, and female investors tend to be more an open book. The open books have low interest in financial matters and have little self-confidence about handling money. While risk-seekers have the most positive attitude toward stocks, the stock market and gambling. Risk seekers would invest higher sums of money in securities. As mostly men are in the 'risk-seekers' group, the men have a more positive attitude towards stocks.

In the study of Wood and Zaichkowsky (2004) 65 percent of their long-term investors cluster, was female. These females have low confidence and control, but do not personalize losses. They trade infrequently. As a group they owned the least number of stocks and do not check their investments often. They purchase long-term conservative mutual funds.

Overall the literature with regards to gender is quite big. Summarizing some outcomes is that men have more trades per year, compared with women. But that men also sell more shorts than women tend to do. But also with regards to men their confidence is bigger than those of women. Due to those facts men are more likely to do more risky trades.

3.1.2 Age

Barber and Odean (2001) found that marital status, age, and income appear to be correlated with the riskiness of the stocks in which a household invests. The young and single hold more volatile portfolios composed of more volatile stocks. They are more willing to accept market risk and to invest in small stocks.

Ameriks and Zeldes (2004) investigated how the household portfolio shares vary with age. They point out that professional financial planners often advise that the fraction of wealth that people should hold in the stock market should decline with age. Although Ameriks and Zeldes (2004) found no evidence of this less holded stocks when aging. In the end they conclude there is no evidence supporting a gradual reduction in portfolio shares with age.

Campbell (2006) suggests that there should be age effects on portfolio choice if older investors have shorter horizons than younger investors and investment opportunities are time-varying, or if older investors have less human wealth relative to financial wealth than younger investors. Campbell (2006) finds in its subgroup of investors in the United States of 2001, that there was a weak negative age effect on participation in public equity markets. This result is presumably due to increased participation by younger households during the 1990s and the fact that the regression controls for wealth and income, which tend to be higher for middle-aged households.

Dorn and Huberman (2005) and Dorn and Sengmueller (2009) found with their subset in Germany in line with Barber and Odean (2001) that younger and male investors trade more aggressively than older and female investors, and also found that older or more experienced and better educated investors hold less concentrated portfolios (Goetzmann and Kumar, 2002). Dorn and Huberman (2009) also found that those who enjoy games only when money is involved, in particular, tend to be younger, less well educated, and less wealthy are more likely to gamble on the stock exchange.

Keller and Siegrist (2006) found that nearly half of the people older than 65 tend to be safe players. Safe players tend to be cautious in financial matters, planning most purchases carefully and large purchases intensively. Safe players also have a negative attitude about stocks, the stock market, and gambling.

Korniotis and Kumar (2009, 2011a) also found evidence that older and more experienced investors hold less risky portfolios, exhibit stronger preference for diversification, trade less frequently, exhibit greater propensity for year-end tax-loss selling.

Lewellen et all. (1977) suggests a narrowing of the return distribution with age-significant at the .0001 level. Thus, the younger investor who engages most heavily in short-run speculation does record the widest range of consequences

In Wood and Zaichkowsky (2004) the confident traders are the oldest group of their long-term investors cluster. These confident traders have the largest investment portfolios, and thus the most experience. They invest heavily in technology and small-cap stocks in their regular portfolios, but maintain a high proportion of stable investments in their retirement portfolios. Older in this case means 97% was older than 30 years.

The literature suggests a lot when it comes to age. It suggests that older people trade less than younger ones do. It also states that older people invest in "less risky" products than younger people.

3.1.3 Wealth

Anderson (2013) found that people with lower levels of wealth and education, and predominantly men, are more prone to stock trading. Anderson also found that wealthier and better educated investors are less sensitive to paper losses, thus assuming they can hold on longer to stocks. Anderson (2013) also found that those with lower levels of wealth and education reduce their trading when their stocks run into losses.

Barber and Odean (2001) also found that the young and wealthy with no dependents are willing to accept more investment risk. Those are thus more willing to invest in small stocks. Campbell (2006) concludes that it appears that poorer and less educated households are more likely to make investment mistakes than wealthier and better educated households.

Calvet, Campbell and Sodini (2008) found some evidence that wealthy, educated investors, hold better diversified portfolios and tend to rebalance more actively. This would suggest that this increases the number of trades.

Ameriks and Zeldes (2004) found that "under a set of simplifying assumptions, a benchmark model of portfolio choice yields the result that the fraction of financial wealth held in the stock market should be independent of both age and wealth. When these assumptions are relaxed, age effects may become important, but there is no uniform prediction about whether the share of wealth held in stocks should increase or decrease with age."

Dorn and Huberman (2005) found that wealthier investors in their sample place more trades, but they turn over their portfolios less frequently, other things equal. However Vissing-Jørgensen (2003) found that wealthier households report placing more trades, via the use of responses from the 1998 and 2001 survey of consumer finances.

Dorn and Sengmueller (2009) found evidence that the male investors and wealthier investors appear to enjoy dealing with investments more than their female and less wealthy counterparts. Those who enjoy games only when money is involved, in particular, tend to be younger, less well educated, and less wealthy.

Grinblatt et al. (2010) found that increases in wealth and trading experiences significantly reduce trading costs. Although, being in the highest wealth quantile reduces trading costs, this only applies for market orders. Grinblatt et al. (2011) found that a statistical decomposition suggests that wealth, income, and education, all influenced by IQ, are key contributors to participation.

Wood and Zaichkowsky (2004) found that of the confident traders in their sample, where confidence is regarded as the ability to invest, more than 50 percent invest more than $ 100.000 and trade more than ten times per year. They also tend to own the most stocks and trade more frequently.

The overall outcome of the literature is that investors trade more often when increasing the level of wealth

3.1.4 Experience and knowledge

Anderson (2013) found suggestions that high-stake investors are less experienced at managing their savings; they have lower wealth and financial wealth, and are younger and less educated. They are more prone to behavioural biases, such as reducing trading when their stocks run into losses.

Barber et al. (2014) investigated the day traders and found that day traders that are experienced and are heavy day traders are more likely to be successful. But both volume and experience are economically weak predictors to past profits. Barber and Odean (2001) report that the differences in self-reported experience by gender are quite large. In general, women report having less investment experience than men.

Dorn and Huberman (2015) found that "investors who think themselves knowledgeable about financial securities indeed hold better diversified portfolios, but those who think themselves more knowledgeable than the average investor churn their portfolios more." Meaning that experienced people mostly trade more.

Grinblatt et all. (2010) mentions that increases in wealth and trading experience significantly reduce trading cost. Korniotis and Kumar (2011a) found evidence that indicates older and more experienced investors hold less risky portfolios, exhibit stronger preference for diversification, trade less frequently, exhibit greater propensity for year-end tax-loss selling. Meaning their choices reflect greater knowledge about investing. But they also found that with the cognitive aging, older investors have worse investment skill, where the skill deteriorates sharply around the age of 70. Collectively Korniotis and Kumar (2011a) found that their evidence indicate that older investors' portfolio choices reflect greater knowledge about investing, but their investment skill deteriorates with age due to the adverse effects of cognitive aging.

Overall the literature suggests that the more experienced traders, tend to trade more, then those who regard themselves as less experienced. Also the literature suggests that people older than 70 years show cognitive aging in their trading cost. Which means that older investors' portfolio choices reflect greater knowledge about investing, but their investment skill deteriorates with age due to the adverse effects of cognitive aging.

3.1.5 Education level(s)

Although some of this literature is already partly treaded, we will summarise the literature found on education.

Anderson (2013) found that individuals that are prone to stock trading on average trade more than others. They have lower levels of wealth and education, and are predominantly male. They are not more successfully when they trade more. Anderson (2013) found also that individuals with lower levels of wealth and education reduce their trading when their stocks run into losses.

Calvet, Campbell and Sodini (2008) found some evidence that wealthy, educated investors, hold better diversified portfolios and tend to rebalance more actively. This would suggest that this increases the number of trades. Campbell (2006) found that it appears that poorer and less educated households are more likely to make mistakes than wealthier and better educated households.

Dorn and Huberman (2005) and Dorn and Sengmueller (2009) found with their subset in Germany also that in line with Goetzmann and Kumar (2002) that older or more experienced and better educated investors hold less concentrated portfolios. Dorn and Sengmueller (2009) also found that those who enjoy games only when money is involved in particular, tend to be younger, less well educated and less wealthy.

Graham et all. (2009) also found that male investors, and investors with larger portfolios or more education, are more likely to perceive themselves as competent than are female investors, and investors with smaller portfolios or less education. They found the relation between the education of an investors and the way they tend to perceive themselves to be more competent, and thus are more willing to act on their beliefs, leading to higher trading frequency.

Keller and Siegrist (2006) found that the majority of the people they regard as open books and risk-seekers have attained higher levels of education than the safe players and money dummies have. With almost 40% having attained a vocational training level (apprenticeship), and about 40% a diploma at the tertiary level (up to the doctorate level).

Korniotis and Kumar (2009a) found that older investors are less effective in applying their investment knowledge and exhibit worse investment skill, especially if they are less educated and earn lower income. Korniotis and Kumar (2009b) also found that the smart investors (better educated, higher income levels and large social networks) who significantly distort their portfolios and hold concentrated portfolios, trade actively, or over-weight local stocks.

Overall one can investigate whether people with lower education invest more in stocks than those with a higher education level. One can also investigate whether people with higher education levels trade more than those with lower education levels do.

3.1.6 Domestic equities

Barberis and Thaler (2003) found that investors exhibit a pronounced "home bias". Investors in the USA, Japan and the UK allocate 94%, 98%, 82% of their overall equity investment, respectively, to domestic equities. Grinblatt and Keloharju (2001) find that investors in that country (Finland) are much more likely to hold and trade stocks of Finnish firms which are located close to them geographically, which use their native tongue in company reports, and whose chief executive shares their cultural background.

Feng and Seaholes (2007) investigated the people's republic of China and found that the degree of home bias is similar across genders – both men and women over-weight local stocks by 9% relative to the market portfolio. Graham et al. (2009) use data from the UBS/Gallup investor survey and found that only 37,5% of all investors hold foreign assets. The remaining 62,5% didn't own any foreign assets. They also found a relation between the competence an investor gave itself and the investment in foreign assets. The higher one says its competence is, the less the home bias gets.

Korniotis and Kumar (2009b) found that recent behavioural literature has shown that individual investors hold concentrated portfolios, trade excessively, and exhibit a preference for local stocks. Meaning that investors invest a disproportionately large proportion of their equity portfolios in geographically proximate stocks. This could be induced by familiarity, where investors over-weight local stocks because they are familiar with them.

Overall one can summaries this section with the suggestion that both gender groups tend to have a very large proportion in local stocks.

3.1.7 Diversification and stake size

Anderson (2013) measured diversification by the investors' stake size, defined as the fraction of their risky financial wealth invested in individual stocks through the broker he studied. High-stake investors have concentrated portfolios, trade more, and achieve lower trading performance. They share several features with those who trade excessively, namely lower income, wealth, age, and education, suggesting that they lack investment expertise. Barber et al. (2009) and (2011) also found that the individuals with no training in investments, hold under-diversified portfolios and so routinely make poor trading decisions.

Barberis and Thaler (2003) found that ambiguity and familiarity offer a simple way of understanding the different examples of insufficient diversification. Investors may find their national stock markets more familiar, or less ambiguous, than foreign stock indices. Feng and Seasholes (2007) also found that both genders are under-diversified and exhibit home bias.

Calvet, Campbell and Sodini (2008) found some evidence that wealthy, educated investors, hold better diversified portfolios and tend to rebalance more actively. Dorn and Huberman (2005) found that the self-reported risk aversion investors tend to diversify most. While more risk tolerant hold less diversified portfolios and trade more aggressively. They again found that less experienced investors tend to churn poorly diversified portfolios. Grinblatt et all. (2011) found that high-IQ investors are more likely to have a higher diversification (from holding mutual funds and greater numbers of stocks).

Korniotis and Kumar (2011a) found evidence that indicates older and more experienced investors hold less risky portfolios, exhibit stronger preference for diversification, trade less frequently, exhibit greater propensity for year-end tax-loss selling. Their choices reflect greater knowledge about investing.

Overall one can summaries this section with the suggestion that less experienced investors hold under-diversified portfolios. Those investors with the willingness towards more risk tend to hold less diversified portfolios and trade more aggressively. While older investors with more experience hold less risky portfolios, with stronger preference for diversification and less trades.

3.1.8 Literature out of scope

A lot of interesting literature can be found with regards to marital status, gambling preference, overconfidence and IQ. In appendix B the literature is worked out. These literature can be a guidance for future research on the subject of cost and investing.

3.1.9 Summary of characteristics and influences

The sub research question of this chapter was:

Is it possible to characterize the different types of investors based on literature research?

Table 1 summarizes the characteristics found during the literature research and summarizes the characteristics per different type of investor. But Table 1 proves it is possible to characterize the different types of investors based on literature research!

Gender
Summarizing some outcomes is that men have more trades per year, compared with women. But that men also sell more shorts than women tend to do. But also with regards to men their confidence is bigger than those of women. Due to those facts men are more likely to do more risky trades.
Age
It suggests that older people trade less than younger ones do. It also states that older people invest in "less risky" products than younger people.
Wealth
The overall outcome of the literature is that investors trade more often when increasing the level of wealth.
Experience and knowledge
Overall the literature suggests that the more experienced traders, tend to trade more, then those who regard themselves as less experienced. Also the literature suggests that people older than 70 years show cognitive aging in their trading cost. Which means that older investors' portfolio choices reflect greater knowledge about investing, but their investment skill deteriorates with age due to the adverse effects of cognitive aging.
Education level
Overall one can investigate whether people with lower education invest more in stocks than those with a higher education level. One can also investigate whether people with higher education levels trade more than those with lower education levels do.
Domestic equities
Overall one can summaries this section with the suggestion that both gender groups tend to have a very large proportion in local stocks.
Diversification and stake size
Overall one can summaries this section with the suggestion that less experienced investors hold under-diversified portfolios. Those investors with the willingness towards more risk tend to hold less diversified portfolios and trade more aggressively. While older investors with more experience hold less risky portfolios, with stronger preference for diversification and less trades.

Table 1 - Summary of characteristics and influences

3.2 Decision tree theory

Decision trees is a method to determine the best predicting factors concerning one dependent variable. The idea of a decision tree is that it creates a tree-based classification model (IBM Corporation [IBM], 2016). It classifies cases into groups or predicts values of a dependent (target) variable based on values of independent (predictor) variables. The tree-based analysis provides some attractive features, as it makes it easy to construct rules for making predictions about individual cases. This description is the idea of the first part of the research question, '*Is it possible to make an estimation of the cost a client has had , based on multiple factors'.*

Four decision tree growing methods are possible, which are:

- Chi-squared Automatic Interaction Detection (CHAID). At each step, CHAID chooses the independent (predictor) variable that has the strongest interaction with the dependent variable. Categories of each predictor are merged if they are not significantly different with respect to the dependent variable
- Exhaustive CHAID. A modification of CHAID that examines all possible splits for each predictor.
- Classification and Regression Trees (CRT). CRT splits the data into segments that are as homogenous as possible with respect to the dependent variable. A terminal node in which all cases have the same value for the dependent variable is a homogenous "pure" node.
- Quick, Unbiased, Efficient Statistical Tree (QUEST). A method that is fast and avoids other methods' bias in favour of predictors with many categories. QUEST can be specified only if the dependent variable is nominal.

Following the description of these four methods, the QUEST method can directly be skipped. This is due to the fact that the transaction cost percentage an investor has is not a nominal variable, but a (continuous) scale variable. This means that the value of the transaction percentage is somewhere between 0.00% and infinity (before winsorizing). The CRT method is not chosen for, as in a CRT model, all splits are binary; that is, each parent node is split into only two child nodes. While in a CHAID model, parent nodes can be split into many child nodes. Concerning the six factors in the dataset, the CHAID applies more to the dataset.

The exhaustive CHAID is a modification to the basic CHAID algorithm, performs a more thorough merging and testing of predictor variables, and hence requires more computing time. Specifically, the merging of categories continues (without reference to any significance level value) until only two categories remain for each predictor. The program then proceeds to choose for the split predictor variable with the smallest adjusted p-value, i.e., the predictor that will yield the most significant split. For the tests the CHAID, as well as the Exhaustive CHAID will be used. The CHAID will be used to get some general insights, after which these insights will be applied to the exhaustive CHAID model.

When conducting the analysis we tick the box with the statistics of the model, which include the summary of the model and the risk, further for the CHAID criteria of splitting nodes, as well as for the merging of categories the significance level is set on 0.05. Further the options boxes 'Adjust significance value using Bonferroni method' and 'Allow resplitting of merged categories within a node' are ticked. Further, due to the fact that both CHAID methods include the missing values in the decision tree, only the investors are selected, with all factors (predictors) filled.

Thus to summarise the steps conducted:

- We use the Chi-squared Automatic Interaction Detection (CHAID) analysis
- The model will summarise the statistics per node including the sample size, mean and standard deviation.

- The output contains the risk estimate and estimation error (more explanation on this output is given in section 4.4.2).
- The CHAID model is allowed to split, but also to merge categories on significance level of 0.05.
- The significance value of the different nodes is adjusted using the Bonferroni method (more explanation with regards to this significance is given in section 5.1).
- The CHAID model is allowed to split a node again after this was already used above as a merge category. This means for example that the wealth categories of €1,000 - €10,000 and 10,000 - €50,000 are first joined together in a node, but are allowed to split up in different nodes further down the tree.
- Only the investors are selected, with a complete data record and thus only those records are used, with all the predictive factors filled in.

3.2.1 Cross validation of decision tree

In section 2.3, the framework of Figure 5 mentions that after finding a conceptual forecasting tool (in the form of a decision tree) we will look for improvements. The use of validation can address this step, as it tests the validity of the decision three. In this section we will work out the theory of the two possible methods of validation. Within SPSS one can test how well the tree structure generalizes to a larger population. SPSS provides two validation method: Crossvalidation and split-sample validation (IBM, 2016).

“Crossvalidation divides the sample into a number of subsamples, or folds. Tree models are then generated, excluding the data from each subsample in turn. The first tree is based on all of the cases except those in the first sample fold, the second tree is based on all of the cases except those in the second sample fold, and so on. For each tree, misclassification risk is estimated by applying the tree to the subsample excluded in generating it.

- We can specify a maximum of 25 sample folds. The higher the value, the fewer the number of cases excluded for each tree model.
- Crossvalidation produces a single, final tree model. The cross validated risk estimate for the final tree is calculated as the average of the risks for all of the trees (IBM, 2016).”

“With split-sample validation, the model is generated using a training sample and tested on a hold-out sample.

- We can specify a training sample size, expressed as a percentage of the total sample size, or a variable that splits the sample into training and testing samples.
- If we use a variable to define training and testing samples, cases with a value of 1 for the variable are assigned to the training sample, and all other cases are assigned to the testing sample. The variable cannot be the dependent variable, weight variable, influence variable, or a forced independent variable.
- We can display results for both the training and testing samples or just the testing sample.
- Split-sample validation should be used with caution on small data files (data files with a small number of cases). Small training sample sizes may yield poor models, since there may not be enough cases in some categories to adequately grow the tree (IBM, 2016).”

In Chapter 5 we will choose between these two methods and explain why that method is chosen.

4 Variables influencing the trading behaviour of clients

In this chapter the following research question will be researched and worked out:

"What variables, influencing trading behaviour of clients, can be found within X?"

At the end of this chapter we should have an adequate dataset. First we need to collect all the necessary data. Then review which factors are out of scope and what factors can be found in the data and might be included in the dataset. The last step is to make all the data applicable for investigation. At the end an adequate dataset is whished for, which is applicable for comparison, including all the determining factors. Afterwards the core statistics and the correlation matrix are given of the complete and adjusted dataset. The chapter ends with the method applied to get to the results.

4.1 Data sources

With the second chapter in mind, the data needs to be applicable to most of the factors and needs to capture most of the cost elements. Due to the fact the data wasn't that easy collected as we thought we had to gather it from three different databases.

4.1.1 Data X clients 'Product'

This was the first data collected, which was collected with the help of the data analyst, working on the department of X 'Y'. In this data the following data was available:

- Account number
- Client number
- Gender
- Age
- Investment amount on 31 March, 30 June, 31 August, 31 December
- X segment (Private Banking, Personal Banking, Mass, etcetera)
- X Sub-Segment (Potential, Youth, etcetera)
- Postal code
- Country of residence
- Aggregated data of transactions
    - Stocks
    - Options
    - Bonds
    - Booster
    - Sprinter
    - Tracker
    - Structured Products
    - Turbo
    - Exchange Traded Fund

These data were separated for each year, so it was divided in the years 2014, 2015, 2016. The original data file had more than 300.000 accounts.

4.1.2 Knowledge and experience

Due to regulation changes in 2014, the data of all new clients and all clients that did a new knowledge and experience (K&E) test, needed to be saved. Due to this fact, it was possible to connect the account number of the clients with these knowledge and experience, or parts of it.

Unfortunately the K&E was changed during the years. So we had to select on a lot of questions and answers and had to combine them later. This is dealt with in section 3.2.2. The query used to collect all the K&E data.

The questions and answers in the K&E we selected to further investigate are those with the topics:

- Education
- Experience due to work
- General experience with investment products

The K&E data collected only had 24.663 clients with a K&E saved in the database, when comparing the account numbers with the data off 4.1.1 (although all clients need to pass the K&E before being allowed to trade, but most investors probably started before 2014, when it wasn't obligatory to save these answers).

4.1.3 Transaction data

Due to the fact that some data of the transactions were missing, it was necessary to load the transaction data of the clients investigated. The only problem was, that more than one million rows of data for the execution only clients, per year, were found. This was too much to even make it applicable for research.

The transaction data was needed, to find out the following:

- Order channel
- Country of Exchange
- Currency
- Investment value
- Category of product

These transaction data is necessary, to be able to compare the transaction data with competitors. This comparison will be used to make a competitor analysis for X. With the use of the K&E a way was found to reduce the huge amount of data and with the inclusion of the K&E in the query, we got the following results.

2014-2015:	237.160 unique transactions
2015-2016:	362.601 unique transactions
2016-2017:	375.181 unique transactions

4.2 Data gathering

In this paragraph we will rephrase the literature and combine this with the data, this summarizes what is worked out. After that we will add some extra variables that might be interesting that popped up when collecting the data.

4.2.1 Match factors with data

In section 3.1.9 Table 1 summarizes the outcome of the literature study and section 4.1 mentions the data sources. It is now possible to select the literature that is left out of scope for future research and what factors will be used for the distinction.

Factors left out

First, considering some outcomes of the gender that are left out is whether men tend to sell short more than women do. What is also left out is whether men feel more confident about trading than women do. Due to the fact a proper way to distinct the short and short sell in the database wasn't available. Also the experience can't be measured and is left out of scope as we decided not to do a questionnaire with the clients of execution only.

The literature that states about risky and less risky trades is left out due to the fact that risky trades and less risky trades are not that straightforward to determine. To determine these, was quite impossible, as more than one million unique transactions were found. So whether men do more risky trades, will be left out of scope.

Whether people older than 70 years show cognitive aging in their trading cost is also left out. This is very difficult to determine, as again in the data, there is no evidence of the trader in the years before. Due to the fact that the dataset is anonymous and thus no unique investor can be distinguished. So whether the 70 years and older show habits of cognitive aging in comparison with those of 60 years and older, will not be worked out.

Overall all of these potential interesting facts to investigate are left out, due to the fact they weren't stored in databases. Above that questioning them via a questionnaire or an interview would disrupt the research and raw data too much. Above that already in section 3.1.8 some distinction was made and some factors were found to be out of scope. These can be found in appendix B.

Insights to investigate out of the data

In the K&E a question is asked whether an investor has experience due to the fact of employment in the field of investing. You might expect that people that work for a bank or broker, would show other investment behaviour than those that don't. A work experienced investor would know how the costs work, what the trick is with buying and selling etcetera. The expectation is that work experienced investors make less cost than those without experience due to work. Or that work experienced investors make less trades per year than those without experience due to work.

Above that a distinction can be made whether orders are made via the call centre or via the internet (as investors pay an extra fee for every order via telephone). One might expect that older people tend to call more than the younger clients of X tend to do, due to the fact that older investors aren't as used to the internet as the younger investors.

Summarizing what variables used for distinction

The factors that will be used to distinct the data are the following:

- Age
- Gender
- Wealth (Amount invested)

- Education
- Work experience
- General experience with investment products

The transaction data found, in combination with the data from X, will be used to distinct the transaction data on different fields, to ultimately find distinction between the cost investors make.

4.2.2 Categorization of variables

Due to the fact that the collected data came from three different databases and weren't universal, a lot of additions were needed. To make clear all the steps made, the next section will state some of those steps and will sometimes point to the appendix to make clear some of the formulas written in Excel.

X clients of product

As described in section 4.1.1 the first dataset started with, was that with generic data and some X segment classes. Above that it included the amount invested and the summed data of transactions per category. Due to some privacy regulations, some information was deleted, to keep the dataset as anonymous as possible.

This resulted in deleting the account number and Client number (after connecting all data on these numbers). The postal code and country of residence was also deleted as these information could be linked to a single person.

The next step was determining the maximum amount invested, due to the fact that some clients had big fluctuations. The decision was to use the maximum value of one of the four time measures (March, June, August, December). After this step the investors were grouped on the amount invested, resulting in six groups:

- Group 1: €0 - €999,99
- Group 2: €1.000 - €9.999,99
- Group 3: €10.000 - €49.999,99
- Group 4: €50.000 – €149.999,99
- Group 5: €150.000 - €499.999,99
- Group 6: €500.000 +

The first group was formed, due to the fact that X had a lot of clients with small amounts of money on their accounts. The sixth group was formed as 500,000 euros is the barrier for the next service fee. The groups in between were formed after some discussions within X and by analysing the dataset.

The same was done with the age of a client, to reduce the amount of output, resulting in seven groups:

- Group 1: 0 - 17
- Group 2: 18 – 25
- Group 3: 26 – 35
- Group 4: 36 – 49
- Group 5: 50 – 64
- Group 6: 65 – 79
- Group 7: 80+

The first group is developed due to the legal age of investing, which is 18 years. If you are younger than this, a (foster) parent or family member older than 18 should deal with the account. The other groups are made after discussion and are based on the use of some common sense. In the Netherlands people can receive a pension around the age of 65, which might result in different trading behaviour before and after reaching this age.

Knowledge and experience data

As mentioned in 4.1.2 the K&E questions changed during the years of analysis. Where first the experience per subcategory was asked it changed to a general question of their experience. Also some rephrases were done, with different numbers meaning the same answers. To sum up what we have done, we will summarize the questions and the answers possible in appendix D.

The first table in appendix E mentions which questions were regarded as the same and how they were made uniform for a comparison. The first column mentions the question number and the second column the specific question corresponding with that number.

The second table in appendix E mentions the answers possible on the selected questions. The third table states which answers were connected with each other. The second column mentions the written and used answer in this research.

After combining the dataset of the K&E with the generic data, the investors that didn't have any question answered were removed. Some investors only answered one off the K&E questions (education, general experience, experience due to work). The questions without an answer, got the answer '0' representing 'no answer'.

Transaction data

The first step in selecting the transaction data was selecting the transactions of execution only per year. But unfortunately, all data, of 2014, 2015 and 2016 were all containing more than one million unique transactions. This was too much to analyse, so after combining it with the K&E, the amount was reduced and the desired result was established.

The second step was checking if all the data was correct, complete and trustworthy. After checking it, some doubtful data was found. For example some funds were categorized as a stock, or another example was that for some stocks, the country of trading was doubtful. Another problem was that the country of stock exchange wasn't available, but was based on their headquarters location. Also the way of categorization of foreign trades was doubtful (i.e. stock bought on AEX-index, placed in Belgium).

This all resulted in the desire to correct them. This formula and some additions, were also used to determine the country of exchange.

The last step was combining the data per account number. Via the use of PowerPivot and Vlookup of Excel, we aggregated the data on some specifications.

4.2.3 Creation of big dataset

For the research the three years were separated with the assumption that (a lot of) differences will exist between the years of 2016, 2015 and 2014. In section 4.3 the core statistics will be described of the three years. When these core statistics show nearly the same mean and standard deviation it might be interesting and better to combine the three years into one big dataset. This big dataset makes it easier and better to select, via the use of SPSS, one or more random sample for crossvalidation. After finding a predictive model, we will apply a cross validation method, to determine the predictive value of the model.

4.3 Descriptive statistics raw dataset

In this subsection the six factors and their frequencies are worked out. As mentioned in section 2.3, the tests are done with the factors tested on the percentage of total transaction cost per year. Then the Winsorize method is worked out and the core statistics, after applying this method, are stated. The sub section ends with the correlation matrix of the factors.

Via the use of one-way analysis of variance (ANOVA) we determine the mean and standard deviation of the factors and the underlying categories. The combined tests, via the use of a decision tree, will be worked out in chapter 5. This section will state the frequencies, median, mean, standard deviation and the minimum and maximum value on a one-dimensional base. Afterwards the correlation of these factors are worked out. All core statistics are from the three years, 2016, 2015, 2014 and these three years combined in one dataset.

4.3.1 Core statistics raw datasets

In appendix F the frequencies, mean, standard deviation, minimal value and maximum value of the years 2016, 2015 and 2014 are mentioned, found by the use of SPSS. The median isn't mentioned due to the fact of the high amount of -% transaction cost accounts. All three datasets have a median of -%

The dataset contains investors that spend the years 2016, 2015 and 2014 with their investment account on hold. This means that these accounts didn't had anything invested in the end of March, June, September and December. These accounts are thus in the dataset with a value of €0 invested in the year 2016, although these accounts pay 'amount' per year as a fee within X they give a not applicable result. This is due to the formula of determining the transaction cost percentage, which is the following:

$$Percentage\ transaction\ cost = \frac{total\ cost\ of\ transactions}{Total\ amount\ invested}$$

All these accounts were changed to the value of zero, as no value on the investment account means that these investors would also have -% transaction cost per year.

All core raw statistics show a high mean and standard deviation. This is supported by the fact that the maximum value for nearly all variables within the factors are more than 'amount' percent. The outcome of these raw core statistics resulted in taking an extra step to deal with the extreme values, which are considered as outliers. Section 4.3.2 will address the different methods of using a barrier or how to exclude some data and section 4.3.3 explains how after these method(s) the outliers were treated.

For example in 2016 a total of 132 accounts had more than 'amount' % transaction cost per year. The biggest invested value of these accounts was €xxx and the average was €xx. 114 investors invested with less than xxxx euros and only 18 invested with more than xxxx.

For 2015 the extreme cases, with accounts higher than 'amount' percent transaction cost, were 126 accounts. The biggest invested value of these accounts was €xxx and the average was €xx. 112 investors invested with less than xxxx euros and only 14 invested with more than xxxx euros.

For 2014 the extreme cases, with accounts higher than 'amount' percent transaction cost, were 77 accounts. The biggest invested value of these accounts was €xxx followed by €xxxxx and the average was €xx. 66 Investors invested with less than xxxx euros and only 11 invested with more than xxxx euros.

The explanation of these extreme value with accounts higher than ‘amount’ percent transaction cost comes most of the times due to the fact of high number of transactions in sprinters, boosters and options, that are done most of the time by an account with a invested amount lower than €x,-. The investor with the highest transaction cost percentage in 2016 for instance had only €xx on its account, but had 2 transactions in boosters and 147 in sprinters. This meant the cost for the boosters was in total €y and for the sprinters this meant €yy (€ z fixed and € zz variable). One can assume that these investors are extreme cases and thus need to be corrected in the datasheet in some sort of matter.

4.3.2 Barriers and other cut-off points

X advises on its webpage a minimal amount to invest with, with regards to investors’ return. This advice is giving due to the fact of the cost. The advice given is to invest a minimal amount of €2,000 for ‘product’. If X advises its new clients and thus new investors this minimal amount it would make sense to make this €2,000 a barrier. To deal with all the zero percent investment cost, this barrier is introduced and also will be used to determine the (best) predictive model.

For the sensitivity analysis, we suggest and will use also a dataset, with a barrier set on €1,000 and on a €100 euros. The last cut-off point used for the sensitivity analysis afterwards is neglecting all investors with a transaction cost percentage of Q%. If investors open an execution only account within X, one might assume that they will invest in the first year. This would mean that Q% transaction cost is not representative for the dataset and can be left out. This assumption makes this dataset it a good starting point for a predictive model.

To conclude we thus will work out the dataset with the barrier set on €2,000. This is done to deal with most off the outliers on the one hand and on the other hand to deal with the zero values. The sensitivity analyses afterwards is done with three different datasets:

- Barrier on €1,000.
- Barrier on €100.
- Neglecting all Q% transaction cost.

4.3.3 Winsorizing or trimming

After setting the barrier at €2,000, we still have to deal with some outliers. The technique of Winsor or the Trimming technique could be applied to the dataset, to deal with these outliers. As Tukey (1977) describes these two techniques to deal with the problem that the tails of a distribution can dominate its value. He describes the Winsor techniques as “one strategy for dealing with this problem is to give less weight to values in the tails of the distribution, and pay more attention to the values near the centre. In essence, winsorizing the distribution changes the highest x% of the scores to the next smallest score, and changes the x% smallest scores to the next largest score.”

Turkey (1977) describes trimming as another strategy for reducing the effects of the tails of a distribution, by simply removing them. “This is the strategy employed by trimming. To find a trimmed mean, the x% largest and smallest scores are deleted and the mean is computed using the remaining scores.”

Due to the fact that nearly 45 percent (10834 neglected investors out of 24328 investors) of the investors in the dataset of 2016 were already neglected, the trimming technique would not be functional, as this results in even neglecting more data. Furthermore the only outliers are found on the upside, as the transaction percentage an investor could have in a year is between 0% and infinity. To cope with the outliers on the upside, the Winsor technique was chosen.

The Winsor technique was then thus applied within SPSS. Appendix G2 states the percentiles for 2016, 2015 and 2014 and combination of three years, per sensitivity analysis point. For winsorizing the percentile points of interest were the 80th, 90th, 95th and 99th percentile. The 80th percentile was inserted as Turkey (1977) uses 20 percent as a measure to winsorize his dataset. The 95th percentile was included as it was used by Gnanadesikan and Kettenring (1972) and Garson (2012). We inserted the 90th percentile as we will only winsorize the upper side outliers and both Gnanadesikan and Kettenring (1972) and Garson (2012) use a winsorize method of ten percent. The value of the 99th percentile was also included as Bali, Engle and Murray (2016) use this in their research. Table 2 summarizes the percentiles and its transaction cost percentage of 2016, 2015 and 2014, but also of the three years combined in one big dataset. These percentages are with the barrier on an invested amount of €2,000 or more. All other winsorize output can be found in Appendix G2, for the dataset of neglecting the Q% transaction percentage, the winsorizing happens on both sides, as the Q values are left out.

Percentage transaction cost								
	20th	25th	50th	75th	80th	90th	95th	99th
2016 (N=13494)	-	-	-	-	-	-	-	-
2015 (N=10580)	-	-	-	-	-	-	-	-
2014 (N=7431)	-	-	-	-	-	-	-	-
Combination of three years (N=31505)	-	-	-	-	-	-	-	-

Table 2 - Output of percentiles with resulting maximum using winsorizing (based on dataset with invested amount ≥ €2,000).

The combination of the three years in Table 2 is used to see whether the values per percentile of the three years combined is in some matter corresponding to the datasets of the years separated. As one can see over the percentiles per year, it is nearly the same for most data and the difference between the three years on the 95th percentile for example is only N% and on the combination the biggest difference is NN%. So we will use the three years combined in one big dataset in the further research.

In section 4.3.4 the winsorized mean and standard deviation of the 95th and 99th percentile is given. In Appendix G3 the SPSS output of the 95th and 99th percentiles, but then per year (2016, 2015 and 2014). In chapter 5 the decision will be made which winsorize method was the best, via the use of the risk estimate of the decision trees. This will be worked out in section 5.1.2.

4.3.4 Core statistics of winsorized dataset

In Appendix H1 the winsorized datasets of 2016, 2015 and 2014 with transaction cost percentage are stated. The winsorized percentiles used and stated are the 95th and the 99th percentile datasets. In Table 3 the core statistics of the winsorized 95th and 99th percentile dataset is stated. This is the output of the big dataset (the three years combined in one dataset). The core statistics contain the mode, mean and standard deviation per factor.

Gender							
	99th percentile				95th percentile		
	N	Median	μ	σ	Median	μ	σ
Male	22924	-	-	-	-	-	-
Female	8581	-	-	-	-	-	-
Total	31505	-	-	-	-	-	-
Experience due to work							
	N	Median	μ	σ	Median	μ	σ
Yes	7366	-	-	-	-	-	-
No	22088	-	-	-	-	-	-
Total	29454	-	-	-	-	-	-
General experience							
	N	Median	μ	σ	Median	μ	σ
No or few experience	7644	-	-	-	-	-	-
Experience	3527	-	-	-	-	-	-
A lot of experience	14807	-	-	-	-	-	-
Total	25978	-	-	-	-	-	-
Education							
	N	Median	μ	σ	Median	μ	σ
Primary education	423	-	-	-	-	-	-
Secondary education	7121	-	-	-	-	-	-
Higher education or university	20969	-	-	-	-	-	-
Other	941	-	-	-	-	-	-
Total	29454	-	-	-	-	-	-
Wealth							
	N	Median	μ	σ	Median	μ	σ
€1000 - €9999,99	12411	-	-	-	-	-	-
€10000 - €49999,99	12738	-	-	-	-	-	-
€50000 - €149999,99	4684	-	-	-	-	-	-
€150000 - €499999,99	1543	-	-	-	-	-	-
€500000 +	129	-	-	-	-	-	-
Total	31505	-	-	-	-	-	-
Age							
	N	Median	μ	σ	Median	μ	σ
0-17	58	-	-	-	-	-	-
18-25	1859	-	-	-	-	-	-
26-35	4115	-	-	-	-	-	-
36-49	7148	-	-	-	-	-	-
50-64	9889	-	-	-	-	-	-
65-79	6597	-	-	-	-	-	-
80 +	1839	-	-	-	-	-	-
Total	31505	-	-	-	-	-	-

Table 3 - Core statistics (99th and 95th percentile) of factors on transaction cost on the combination of three years

4.3.5 Correlation matrix of factors

The next step is testing the correlation between the factors, with the controlling variable (within SPSS) being the winsorized 95th percentile transaction cost. This means it takes into account the influence of the transaction cost. Not surprising is the fact that a lot of correlation values are found to be significant. In appendix H2 the complete correlation figure of the 99th percentile is shown and Table 5 shows the correlation and significance of the 95th percentile dataset.

The test conducted is the two tailed partial correlation tested on the six factors and the transaction percentage cost. The Pearson correlation reviews the linear connection between two factors. To address the SPSS output even further, we will shortly review the factors. Due to the fact that these factors are used for distinction and no intrinsic value is seen in the different answers given by the investor. The values of the different factors are thus treated as a nominal value, thus the values represent categories with no intrinsic value. The percentage cost is a scale variable, which means that the values represent (ordered) categories with a meaningful metric. In our case, the output of the transaction cost in percentage is the (ordered) categories on a meaningful metric.

If the correlation is found to be significant, this means that the correlation is not zero. A nonzero correlation could exist, meaning that an association might be found. If no significance is found, it withholds that on the significance levels tested, the correlation coefficient is zero and there is thus no association found.

We will review the correlation found to be significant and with an higher correlation value than 0.30 (r-value ≥ 0.30). This is due to the fact that the square of this r-value addresses the proportion of explained variance between the two factors. Taking a value of r > 0.30 means we address all correlations with an proportion of explained variance higher than nine percent.

When factors have correlation values between 0.3 and 0.5 they are considered as weakly correlated, if factors have correlation values between 0.5 and 0.7 they are considered as moderate correlated. If factors have correlation values between 0.7 and 0.85 they are considered as strongly correlated, if factors have correlation values between 0.85 and 0.95 they are considered as very strong correlated. Factors with a correlation value above the 0.95 are considered as extremely strong correlated (Doorn and Rhebergen, 1998).

	GENDER	EXPERIENCE DUE TO WORK	GENERAL EXPERIENCE	EDUCATION	WEALTH	AGE	PERCENTAGE COST
GENDER	1.000	0.062***	-0.005	0.059***	0.050***	0.096***	-0.115***
EXPERIENCE DUE TO WORK	0.062***	1.000	-0.099***	**0.999*****	-0.054***	-0.016**	-0.059***
GENERAL EXPERIENCE	-0.005	-0.099***	1.000	-0.102***	0.071***	0.048***	-0.019***
EDUCATION	0.059***	**0.999*****	-0.102***	1.000	-0.053***	-0.023***	-0.061***
WEALTH	0.050***	-0.054***	0.071***	-0.053***	1.000	**0.325*****	-0.191***
AGE	0.096***	-0.016**	0.048***	-0.023***	**0.325*****	1.000	-0.142***
PERCENTAGE COST	-0.115***	-0.059***	-0.019***	-0.061***	-0.191***	-0.142***	1.000

Table 4 - Correlation value and significance (p < 0.05. **p < 0.01. ***p < 0.001.) of factors and percentage transaction cost*

An extremely strong correlation is found between education and experience due to work. The correlation value of 0.999 suggests that 99.80 percent of the variance between these two factors is explained by the other factor. Further a weak positive correlation is found between the age factor and the wealth factor, with a value of 0.325. Suggesting that 10.56 percent of the variance between these two factors is explained by the other factor.

4.4 Risk estimates output

Decision trees is a method to determine the best predicting factors concerning one dependent variable (IBM, 2016). In this sub section we determine the estimation error and thus the estimation power of the dataset, on four different Winsor (99th -, 95th, 90th and 80th percentile) levels.

With decision trees it is also possible to target the risk estimate and the estimation error (IBM, 2016). One indicator of the model's performance is the risk estimate. The percentage of transaction cost is a scale dependent variable in SPSS (IBM, 2016). As the transaction cost percentage is continuous between 0.00% and infinity (before winsorizing). "For a scale dependent variable, the risk estimate is a measure of the within-node variance, which by itself may not tell you a great deal. A lower variance indicates a better model, but the variance is relative to the unit of measurement. If, for example, price was recorded in ones instead of thousands, the risk estimate would be a thousand times larger."

"To provide a meaningful interpretation for the risk estimate with a scale dependent variable requires a little work:

- Total variance equals the within-node (error) variance plus the between-node (explained) variance.
- The within-node variance is the risk estimate value, that SPSS provide with the Risk output of the decision tree.
- The total variance is the variance for the dependent variables before consideration of any independent variables, which is the variance at the root node.
- The standard deviation is displayed at the root node; so the total variance is that value squared (IBM, 2016)."
- The proportion of variance due to error (unexplained variance) is:

$$Unexplained\ variance = \frac{Risk\ estimate}{(Standard\ deviation\ at\ root\ node)^2}$$

- The proportion of variance explained by the model is thus:

$$Explained\ variance = 1 - Unexplained\ variance$$

With the use of the description provided, it is possible to determine the predictive value of the variance, by the use of the risk estimator of the decision trees. In Table 5 the predictive value of the four datasets adjusted by the Winsor method is stated for the big dataset, with a barrier on the amount invested on €2000. Table 6 gives that of the big dataset, where the transaction cost percentage is bigger than Q%. The other two barrier datasets are placed in Appendix H3.

(Barrier: Invested amount ≥ €2,000)	Risk estimate	Standard error	Standard deviation at root node	Unexplained variance	Explained variance
99th percentile winsorized dataset	-	-	-	96.19 %	3.81 %
95th percentile winsorized dataset	-	-	-	93.69 %	6.31 %
90th percentile winsorized dataset	-	-	-	92.30 %	7.70 %
80th percentile winsorized dataset	-	-	-	91.16 %	8.84 %

Table 5 - Winsorized predictive value of decision trees

(Transaction cost percentage of all non-zero)	Risk estimate	Standard error	Standard deviation at root node	Unexplained variance	Explained variance
Original dataset	-	-	-	99.99%	0.01%
99th percentile winsorized dataset	-	-	-	90.70 %	9.30 %
95th percentile winsorized dataset	-	-	-	79.11 %	20.89 %
90th percentile winsorized dataset	-	-	-	73.79 %	26.21 %
80th percentile winsorized dataset	-	-	-	68.33 %	31.67 %

Table 6 - Winsorized predictive value of decision trees (dataset: transaction cost percentage is non zero)

The output of Table 6 suggests that the use of the combined dataset of neglecting the transaction cost percentage, is higher on all winsorize levels compared with the dataset we used in section 4.3 (invested amount ≥ €2,000). We will thus from now on continue with using the transaction cost dataset (transaction cost > Q%). The barriers are still kept for a sensitivity analyse.

The three barriers used are thus:

- Invested amount ≥ €2,000.
- Invested amount ≥ €1,000.
- Invested amount ≥ €100.

The winsorized predictive value of these decision trees are placed in Appendix H4. All barriers seem to have less predictive power than only neglecting the Q percent values. The original dataset, the one neglecting the Q% but without any barrier, is thus used to determine the best predictive model.

4.5 Summarizing the sub research question

To summaries what is done this chapter, we will review the sub research question, answer it and briefly summarizes the dataset used and its core statistics. First we rephrase the sub research question.

“What variables, influencing trading behaviour of clients, can be found within X?“

The factors, influencing trading behaviour of clients, found within X after combining and adjusting the dataset, are: Age, gender, wealth (amount invested), education, work experience and general experience with investment products. These factors themselves are sub divided in categories, like for example the factor ‘general experience with investment products’ is divided in these three categories: ‘No or few experience’, ‘experience’ and ‘A lot of experience’.

After determining the factors to test on, we investigated the core statistics and found two problems, one problem is the transaction cost percentage and the other problem is that on the upperside of the transaction cost percentage, some very extreme outliers were found. With the use of four different barriers, these problems were dealt with. The four different barriers were: neglect all Q transaction cost percentage, invested amount ≥ €100, invested amount ≥ €1000 and invested amount ≥ €2000.

After winsorizing on four different levels, that of the 99th percentile, 95th percentile, 90th percentile and the 80th percentile, the best predictive dataset was found, with the use of the formula to determine the explained variance. In Chapter 5 the dataset of neglecting the Q transaction cost percentage will be worked out, as Table 6 shows that this dataset explains the most variance.

5 Results

In this chapter the following research question will be researched and worked out:

Is it possible to make an estimation of the cost a client has had , based on multiple factors and are these cost useful to make a yearly prediction of the expected costs?

To get to an estimation of the cost an investor has had, based on multiple factors, the decision tree function of SPSS will be applied. The decision tree we use and work out, is the 90th percentile winsorized mean dataset. The 90th percentile is chosen as it has 5.42 percent more explained variance than the 95th percentile had. Choosing the 80th percentile would have resulted in even more explained variance (increase of 5.46 percent), but also means we have to reduce the raw dataset with even 10 percent more winsorizing.

This tree is worked out in detail and after finding the predicting decision tree, we will review the strength of the decision tree via the cross validation function. Also the strength of the three sensitivity analysis trees will be worked out, to see if the best possible model is used. Their predictive power is placed in the Appendix. After the sensitivity analysis, a review and explanation is done on how to show and calculate the cost, including the interval per node.

5.1 Decision tree

To get the best predictive decision tree, the decision tree was tested on three significance levels for the splitting nodes and the merging categories in the criteria tab of SPSS. In this subsection we will first only highlight the outcome of the three resulting trees. As mentioned in section 2.3 in the methodological framework, due to simplicity reasons, the model will be bound to a maximum of 3 to 4 variables. This first step is thus done, to indicate which factors are included in the decision tree. The best predictive decision tree, will be decided via the explained variance method in section 5.1.2.

By first checking the models, with a significance level of 99 percent and 95 percent, on their first node(s), second node(s) and third node(s), the order of testing will be determined. The CHAID model is allowed to split, but also to merge categories on significance level of 0.01 and 0.05. Afterwards, by combining these factors, the best possible decision tree will be selected on the predictive value method as before. Afterwards tests will be conducted, if adding a fourth factor increases the predictive value of the decision tree. The dataset with only transaction cost percentage higher than Q % is worked out in section 5, the predictive power of the other three datasets are mentioned in Appendix I1.

The significance level we test on in section 5.1.2 and further is the confidence level of 95%. This means that we thus test the (null) hypothesis that no difference exists between the different nodes. If a p-value lower than 0.05 is found we reject the (null) hypothesis and thus reject the assumption that no difference between the different nodes exists. We will then assume with 95 percent certainty that indeed a 'significant' difference exists between the nodes.

5.1.1 Outcomes tree nodes

This first sub section is worked out, to briefly introduce the decision trees found within SPSS. No output will be placed in figures or tables, due to the fact that this sub section is purely worked out to indicate which factors are found when using the winsorized 90th percentile dataset (dataset used is the transaction cost higher than Q percent). Also on the three other datasets (barrier on invested amount of 2000, 1000 and 100 euros) tests are conducted to see which factors are included in those datasets and their decision trees. Only if a differences occur when comparing the factors included in the main decision tree and those decision trees, the differences and the different factors will be mentioned.

The first test conducted was on a significance level of 99% for both the merging of categories and the splitting of nodes. The First node is build up by the wealth factor. In total 5 nodes are on the first row of nodes. On the second row of the decision tree the following factors were found, one time the node is divided by gender, two times by education and two times by the general experience of an investors. On the third row the factors that were found in dividing the investors was four times the gender factor and happens two times on education.

The second decision tree made and conducted was on a significance level of 95% for both the merging of categories and the splitting of nodes. Again the first node is build up by the wealth factor. In total again 5 nodes are created. The second row of the tree is again divided by one time the gender factor, two times the education factor and two times the general experience of an investor. The third row is now built up by five times the gender factor and three times the education factor.

The three datasets (barrier on invested amount of 2000, 1000 and 100 euros) left for the sensitivity analysis show nearly the same decision tree, as the first two rows of the decision tree correspond one on one and on the third row level, only one time general experience is chosen as factor instead of the education variable. But above that small detail, also in these decision trees the factors that are only included are: ‘Wealth’, ‘education level’, ‘gender’ and ‘general experience’. The analysis of the decision trees of the four different datasets show that the added value of the factors age, as that of experience due to work are not that substantial in their added value. In the section 5.1.2 we will use the four factors left and those four factors will be combined to check the decision tree and eventually their predictive value combined.

5.1.2 Risk estimates four factors left

Again the predictive value of the factors will be determined, with the use of the winsorized 90th percentile dataset. This will be done in the same matter as happened in section 4.4.2. This means that the decision tree nodes will be developed on a significance level of 95 percent. First the combined CHAID output on three levels with all four factors included is tested. The second row is the output of the four factor on a decision tree with four levels. The third row is the factors wealth, gender and education combined in a decision tree. The fourth row is the combination of the factors wealth gender and general experience. The fifth and last row is the combination of the factors wealth, education and general experience. The results of the predictive value is found in Table 7. For the three datasets the predictive value tables are placed in Appendix I1.

Measurement	Risk estimate	Standard error	Standard deviation at root node	Unexplained variance	Explained variance
Four factors, three levels decision tree	-	-	-	73.79%	26.21%
Four factors, four levels decision tree	-	-	-	73.54%	26.46%
Wealth, gender and education	-	-	-	73.03%	26.97%
Wealth, gender and general experience	-	-	-	74.36%	25.64%
Wealth, education and general experience	-	-	-	74.09%	25.91%

Table 7 - Predictive value of decision tree on different factors, for dataset with transaction cost > Q percent.

The rank of the explained variance in Table 7 is thus that factors wealth, gender and education combined have a predictive value of 26.97%, second place is the four factors combined on a four levels decision tree, with a predictive value of 26.46%. Third best is the combination of the four factors on a three levels decision tree. This means that the factor of general experience can better be left out of a decision tree. The decision trees where the factor general experience is introduced, the explained variance is found to be less than when we applied the four factors on a three levels decision tree. We will thus work out the decision tree of wealth gender and education in detail in section 5.1.3.

To be really sure of selecting the best possible decision tree, under the assumption of having the most variance explained, we also tested the predictive value of datasets with two assumptions combined. This means that the used dataset (transaction cost percentage higher than Q percent), is combined with the barriers of invested amount. The winsorized 90th percentile is also used for these tests. The predictive value is also not higher than the combination of wealth, gender and education in Table 7. We will thus continue using the factors wealth, gender and education in section 5.1.3. The decision tree of the three factors combined will be worked out on a significance level of 95% for both the merging of categories and the splitting of nodes. This will be done with the CHAID method.

5.1.3 Decision tree

As mentioned the decision tree of the factors wealth, gender and education will be worked out in this section. This will be done on a significance level of 95% for both the merging of categories as for the splitting of nodes. The tree is developed with the CHAID method.. Figure 6 summarizes the decision tree and Table 8 gives the summary output of the decision tree.

The significance level we test the output on is on the confidence level of 95%, 99% and 99.9%. This means that we thus test the (null) hypothesis that no difference exists between the different nodes. If a p-value lower than 0.05, 0.01 or 0.001 is found we reject the (null) hypothesis and thus reject the assumption that no difference between the different nodes exists. We will then assume with 95, 99 or 99.9 percent certainty that indeed a 'significant' difference exists between the nodes.

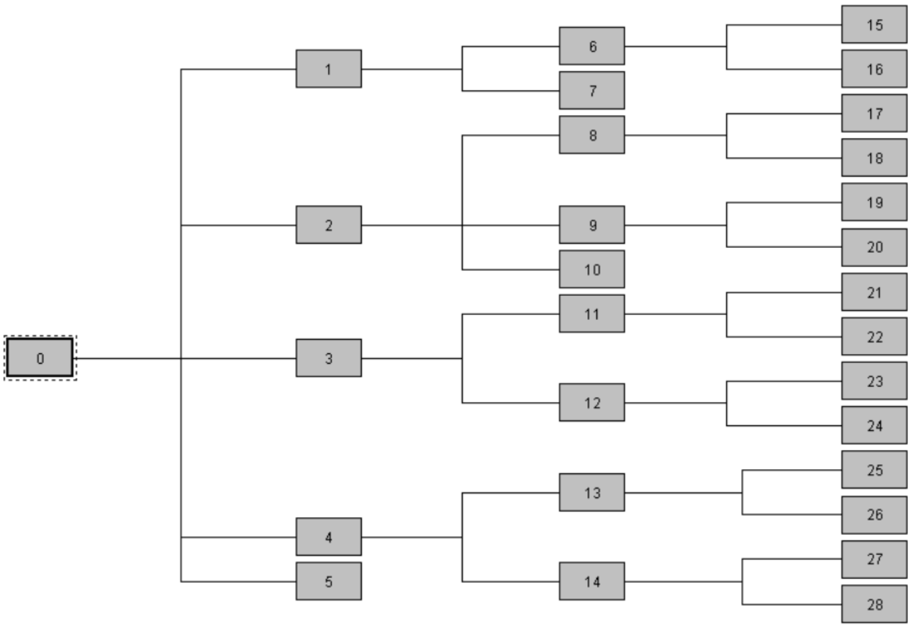

Figure 2 - Decision tree based on the factors wealth, gender, age and education

Node	μ	σ	N	Percent	Parent Node	Variable	Sig.	Split Values
0	-	-	22092	-				
1	-	-	3737	-	0	wealth.cat	0.000***	€0 - €999,99
2	-	-	8789	-	0	wealth.cat	0.000***	€1.000 - €9.999,99
3	-	-	6419	-	0	wealth.cat	0.000***	€10.000 - €49.999,99
4	-	-	2357	-	0	wealth.cat	0.000***	€50.000 - €149.999,99
5	-	-	790	-	0	wealth.cat	0.000***	€150.000 - €499.999,99; €500.000+
6	-	-	3087	-	1	Gender	0.000***	Male
7	-	-	650	-	1	Gender	0.000***	Female
8	-	-	2486	-	2	Education.lev	0.000***	Other; Secundary Education
9	-	-	6215	-	2	Education.lev	0.000***	Higher Education or University
10	-	-	88	-	2	Education.lev	0.000***	Primary Education

11	-	-	1920	-	3	Education.lev	0.000***	Other;
								Secundary Education;
								Primary Education

12	-	-	4499	-	3	Education.lev	0.000***	Higher Education or University

13	-	-	1749	-	4	Education.lev	0.000***	Other;
								Higher Education or University;
								Primary Education

14	-	-	608	-	4	Education.lev	0.000***	Secundary Education
15	-	-	2218	-	6	Education.lev	0.000***	Other; Higher Education or University
16	-	-	869	-	6	Education.lev	0.000***	Secundary Education; Primary Education
17	-	-	1971	-	8	Gender	0.000***	Male
18	-	-	515	-	8	Gender	0.000***	Female
19	-	-	5027	-	9	Gender	0.000***	Male
20	-	-	1188	-	9	Gender	0.000***	Female
21	-	-	1479	-	11	Gender	0.000***	Male
22	-	-	441	-	11	Gender	0.000***	Female
23	-	-	3478	-	12	Gender	0.000***	Male
24	-	-	1021	-	12	Gender	0.000***	Female
25	-	-	1313	-	13	Gender	0.010*	Male
26	-	-	436	-	13	Gender	0.010*	Female
27	-	-	448	-	14	Gender	0.033*	Male
28	-	-	160	-	14	Gender	0.033*	Female

Table 8 - Summary of decision tree, corresponding with Figure 6 (p < 0.05. **p < 0.01. ***p < 0.001.)*

As mentioned in section 5.1.2 in Table 7 the factors wealth, education and gender combined have a predictive value of 26.97%. In section 5.2 the three factors are combined in a decision tree on the same matter, but now with the use of cross validation to determine more precise the predictive value of the tree.

5.2 Review decision tree with the use of cross validation.

After determining the explained variance and using this variance to decide which combination of factors is the best according to this method, we need to review the model itself. This will be done in this subsection. In subsection 3.2.1 the explanation of the two different methods for crossvalidation are worked out. In this section we state the output of the cross validation and thus the predictive value of the decision tree. We hope that after cross validating the decision tree, we get the same results and find the decision tree as reliable as in subsection 5.1.

When reviewing the two different approaches to validate, we have chosen for the crossvalidation method. Due to the fact that this method produces one final decision tree, while reviewing 25 (random) sample folds. Within SPSS we choose for 25 sample folds (the maximum folds possible to select) and as a result we got the final tree model, including the risk estimate. All output is worked out in subsection 5.2.1.

5.2.1 Crossvalidation

As mentioned in section 5.2 the validation method used, to review the decision tree, is the 'crossvalidation' function of SPSS. After getting a cross validated risk estimate, we were able to determine the explained variance of the model after validation. The output within SPSS resulted in the same outcomes as we found in Table 8 and Figure 6. Although Table 9 states the output of the cross-validation risk estimate of the decision tree and the complete SPSS output is placed in Appendix J1.

As can be seen the decision tree explains 26.77 percent of the variance, making the reliability of the predictive model questionable. As only one fourth of the variance between the factors is explained. This indicates that much of the variation in the data remains unexplained by the model (Gelman and Pardoe, 2006). Gelman and Pardoe (2006) also conclude that one fourth of the variance explained is low. But due to the fact that this is the best possibility, we continue with working out the output predicted by this validation model in section 5.3.

Measurement	Crossvalidation risk estimate	Standard error	Standard deviation at root node	Unexplained variance	Explained variance
Wealth, gender and education Crossvalidation	-	-	-	73.23%	26.77%

Table 9 - Crossvalidation risk estimate output and explained variance complete model.

5.3 Forecasting the cost per node

After determining the reliability and predictive value of the model and reviewing it via the crossvalidation, the only step left is determining and working out the model. Although we will not address the topic about the design or in what way X should ask the questions to determine the new investors its transaction cost per year. However we will suggest what kind of content X could show on their webpage.

First of all determining the mean and standard deviation is not enough. What we suggest is showing the customer, an certain interval in which his cost probably will be, with a significance level of 95 percent. What we would suggest is showing the 5th percentile as well as the 95th percentile of all the investors in the node, above that we suggest to give the average or mean to the customer. In Table 10 we mention also the standard deviation and also the 2.5th and 97.5th percentile, to give X the choice which interval they prefer to show to new customers (or they can show both). Figure 7 is shown to know where in the node the output is given.

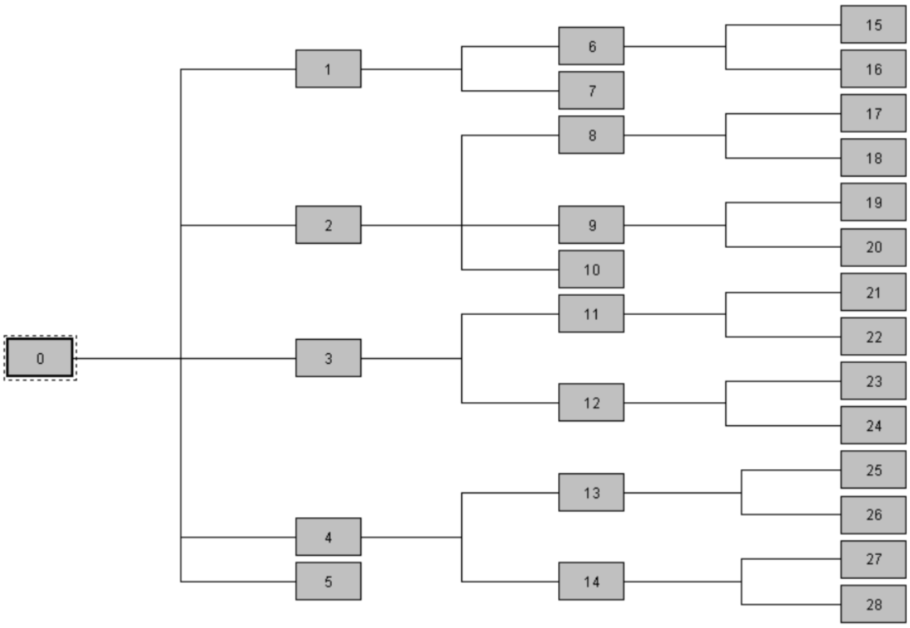

Figure 3 - Decision tree based on the factors wealth, gender, age and education

Node	μ	σ	N	Median	2.5th perc.	5th perc.	95th perc.	97.5th perc.	Split Values
0	-	-	22092	-	-	-	-	-	
1	-	-	3737	-	-	-	-	-	€0 - €999,99
2	-	-	8789	-	-	-	-	-	€1.000 - €9.999,99
3	-	-	6419	-	-	-	-	-	€10.000 - €49.999,99
4	-	-	2357	-	-	-	-	-	€50.000 - €149.999,99
5	-	-	790	-	-	-	-	-	€150.000 - €499.999,99; €500.000+
6	-	-	3087	-	-	-	-	-	Male
7	-	-	650	-	-	-	-	-	Female
8	-	-	2486	-	-	-	-	-	Other; Secundary Education
9	-	-	6215	-	-	-	-	-	Higher Education or University
10	-	-	88	-	-	-	-	-	Primary Education
11	-	-	1920	-	-	-	-	-	Other; Secundary

									Education; Primary Education
12	-	-	4499	-	-	-	-	-	Higher Education or University
13	-	-	1749	-	-	-	-	-	Other; Higher Education or University; Primary Education
14	-	-	608	-	-	-	-	-	Secundary Education
15	-	-	2218	-	-	-	-	-	Other; Higher Education or University
16	-	-	869	-	-	-	-	-	Secundary Education; Primary Education
17	-	-	1971	-	-	-	-	-	Male
18	-	-	515	-	-	-	-	-	Female
19	-	-	5027	-	-	-	-	-	Male
20	-	-	1188	-	-	-	-	-	Female
21	-	-	1479	-	-	-	-	-	Male
22	-	-	441	-	-	-	-	-	Female
23	-	-	3478	-	-	-	-	-	Male
24	-	-	1021	-	-	-	-	-	Female
25	-	-	1313	-	-	-	-	-	Male
26	-	-	436	-	-	-	-	-	Female
27	-	-	448	-	-	-	-	-	Male
28	-	-	160	-	-	-	-	-	Female

Table 10 - Final output of model, including mean, standard deviation, median and two interval levels.

5.3.1 Examples of output

In this sub section we will work out two examples, of two different new clients and what we suggest to give as output. The examples we will work out are the following two investors.

1. Male investor, wanting to invest an amount of €12,345 and the investor his highest education level is secondary education.
2. Female investor, wanting to invest an amount of €120,000 and the investor her highest education is (higher education or) university.

For the first investor we will first review the table. The amount he is willing to invest, is more than €10,000 but less than €50,000 and thus he belongs to the second node. His highest education level is secondary education and thus he will continue to node 8. As the investor also filled in his gender, we can finalize and find the output in node 17. The percentage corresponding to that node are the following.

Node	µ	σ	N	Median	2.5th perc.	5th perc.	95th perc.	97.5th perc.
17	-	-	1971	-	-	-	-	-

Table 11 - Values corresponding with profile investor 1

Before suggesting what we think could be interesting as output, we will rephrase the equations to give the investor an idea of the percentages but also to give them insights in the gross cost per year

$$\begin{aligned} predicted\ average\ cost &= transaction\ cost + service\ fee \\ predicted\ median\ cost &= transaction\ cost + service\ fee \\ predicted\ lower\ interval\ cost\ (5\%) &= transaction\ cost + service\ fee \\ predicted\ higher\ interval\ cost\ (95\%) &= transaction\ cost + service\ fee \end{aligned}$$

X could now for instance correspond to the investor. Based on your profile, we found 1971 investors, who on average paid x% per year as transaction cost. Including the service fee your cost per year are the following. First we show how and what the service fee is and then this is simply added on the calculated gross transaction cost. The investor has a service fee of 'a' percent when it invests with less than €z and between €z and €zz this is 'aa' percent.

$$\begin{aligned} service\ fee\ investor\ 1 &= (€\ 12345 * a\%) + €a = €\ aa \\ predicted\ average\ cost &= (€\ 12345 * b\%) + aa = €\ bb \\ predicted\ median\ cost &= (€\ 12345 * c\%) + aa = €\ cc \\ predicted\ lower\ interval\ cost\ (5\%) &= (€\ 12345 * d\%) + aa = €\ dd \\ predicted\ higher\ interval\ cost\ (95\%) &= (€\ 12345 * e\%) + aa = €\ ee \end{aligned}$$

We also suggest to give the investor the complete cost per year also in a percentage and thus apply the following formulas:

$$\begin{aligned} service\ fee\ investor\ 1 &= €\frac{aa}{12345} * 100\% = aaa\% \\ predicted\ average\ cost &= €\frac{bb}{12345} * 100\% = bbb\% \\ predicted\ median\ cost &= €\frac{cc}{12345} * 100\% = ccc\% \\ predicted\ lower\ interval\ cost\ (5\%) &= €\frac{dd}{12345} * 100\% = ddd\% \\ predicted\ higher\ interval\ cost\ (95\%) &= €\frac{ee}{12345} * 100\% = eee\% \end{aligned}$$

So to summaries we would suggest X to mention the cost in gross cost per year, but to also mention what this is completely per year as a percentage. We although suggest to don't show any calculations on how the values are determined, but to make a button, with something like "interested in how this is determined". On this page the investor could see the limitations an assumptions applied in this research and that the model is based on real investors within X, who were monitored for a period of three year.

To briefly address how to deal with the calculation of the cost of the female second investor, via the same way as done with the first investor, we end in node 26:

Node	µ	σ	N	Median	2.5th perc.	5th perc.	95th perc.	97.5th perc.
26	-	-	436	-	-	-	-	-

Table 12 - Values corresponding with profile investor 2

$$service\ fee\ investor\ 2 = (€\ 75000 * a\%) + (45000 * aa\%) + €a = €\ aa$$

$$predicted\ average\ cost = (€\ 120000 * b\%) + aa = €\ bb$$
$$predicted\ median\ cost = (€\ 120000 * c\%) + aa = €\ cc$$
$$predicted\ lower\ interval\ cost\ (5\%) = (€\ 120000 * d\%) + aa = €\ dd$$
$$predicted\ higher\ interval\ cost\ (95\%) = (€\ 120000 * e\%) + aa = €\ ee$$

$$service\ fee\ investor\ 2 = €\frac{aa}{120000} * 100\% = aaa\%$$
$$predicted\ average\ cost = €\frac{bb}{120000} * 100\% = \ bbb\%$$
$$predicted\ median\ cost = €\frac{cc}{120000} * 100\% = \ ccc\%$$
$$predicted\ lower\ interval\ cost\ (5\%) = €\frac{dd}{120000} * 100\% = \ ddd\%$$
$$predicted\ higher\ interval\ cost\ (95\%) = €\frac{ee}{120000} * 100\% = \ eee\%$$

To wrap it up, we suggest X to tell the new investor, after this investor filled in the required field about their investment profile. The factors will lead to the corresponding node found within SPSS. Then in the background the amount the investor is willing to invest is used to calculate the output what this investor would probably pay. We suggest to give the investor the choice to see a percentage of gross cost, but it is also possible to show both.

The output could thus be: “Based on your investment profile, we found 436 corresponding investors. Based on their profile they would have an average yearly cost of € ‘a’, or ‘a’ percent. 90 percent of the investors have a yearly cost between the €b and €c, or b percent and c percent and based on the median 436 investors the median investor paid €d, or d percent per year.” What might be interesting is to show them a bar in which all variables are shown. But we also suggest some sort of table like table 13, that summarizes what is stated.

Mrs. Groot	Gross yearly cost	Percentage per year
Average yearly cost	€ bb	bbb%
Lower interval cost (5th percentile)	€ cc	ccc%
Median (50th percentile)	€ dd	ddd%
Upper interval cost (5th percentile)	€ ee	eee%
Based on 436 investors, Xs’ prediction of your yearly cost is.		

Table 13 - Summary of investor 2, mrs. Groot

We suggest to put underneath the output directly a link towards the investigation and the assumptions and limitations of the investigation. This would be in the form of something like: “Interested in how this is determined, click here”.

Further the standard notes of X need to be added, like ‘investing is not without risks, one could lose all invested amount.’ But also something like, ‘history isn’t a guarantee for the future, this predication is on an average base and based on a dataset from X, the real cost per year could be different than what is predicted’. To conclude with something like ‘X doesn’t guarantee any transaction cost per year, the model is purely to give an indication and no legality can be granted from the model.’

5.4 Answering sub research question

In this chapter the following research question was researched and worked out:

Is it possible to make an estimation of the cost a client has had , based on multiple factors and are these cost useful to make a yearly prediction of the expected costs?

Concluding the possibility to make an estimation is there, as seen in section 5.3. But as the predictive value of the decision tree is found to be 26.77%, after conducting a crossvalidation research. As only one fourth of the variance between the factors is explained, making the reliability of the predictive model questionable. This indicates that much of the variation in the data remains unexplained by the model, making it not applicable for publication on webpages and applications. Nevertheless the decision tree gave some good insights in, the best predictive factors, the way they can be distributed and divided, and the distinction between the nodes. The chapter ends with two examples and on how the content might be showed by X on the webpage.

To summarize and answer the research question, it is possible to make an estimation of the cost a client has had, based on three factors (wealth, gender and education) and the corresponding values out of the (decision tree) model are useful to make a yearly prediction of the expected cost. But the predictive power is found to be low, with only one fourth of the variance explained. Overall the model gives nice insights in the investors cost, but X should really question themselves if it is smart to publish it on their webpage and in their new application. Our advice is to don't publish it, but to use the output for targeting some groups. We will work out our conclusion in detail in Chapter 7.

7 Conclusion, limitation and further research

The problem statement of this whole report is about *"How could X increase its user friendliness with regards to the information on costs of investments for investors, keeping in mind regulatory needs and how should X show these costs omnichannel?"* we will first give an answer to this problem statement, will than continue with the limitations of this research and will and with suggestions for further research.

7.1 Conclusion

After investigating the literature with the suggestions that some differences appear between investors, for example compared on gender, wealth age, experience and education. After collecting the dataset and discussing what was desired, a decision tree model was made. The idea of this predictive model was to give new investors arriving at the execution page of X (website and application on mobile device), a change to get insight in their yearly cost based on their characteristics.

After a lot of calculations, adjustments and assumptions, we should although say that some characteristic differences occur, but that the combination of wealth, gender and education isn't as predictive as one might have hoped. The predictive value (explained variance) is 26.77 percent, after validating the dataset. Although this combination was found to be the best and although a lot of alternatives were thought off, no combination or dataset outperformed this dataset in predictive value. Overall we should conclude that making a predictive model for new investors, interested in execution only within X, isn't as trustworthy as one might have thought it would be.

Although we would suggest X to continue the investigation on these characteristics in the future. Although the decision tree and its predictive value isn't as trustworthy, the dataset and SPSS output provided some insights on the literature used. In the dataset provided by X, based on the transaction cost paid per year, significant differences (95 percent confidence interval) where found between gender, education level, experience in investing, wealth and age.

Male investors tend to pay more transaction cost per year than female investors do. Investors with a lot of experience are paying significantly less than those with no or a few experience. The cost per year paid, based on the education level is that investor with a lower education level tend to pay more than investors with a higher education level. The distinction made is here between primary, secondary and university (or higher) education. Also higher levels of wealth are significant different than the ones with less wealth. This suggests, as the literature also stated, that increases in wealth also increases the number or at least the value of the transactions done. Although the transaction cost as a percentage decline per higher wealth group. When looking into age, we saw a decline per age group. Being the highest for the persons of 18 years to 25 years. The cheapest transaction cost per year were for those investors older than 80 years.

7.2 Limitations

The idea of the research was to give the new investors, with interest in Xs execution only, a prediction of the cost they could expect per year. After some calculation and connections in Excel a complete dataset was established. But in the connections and calculation, although it was checked twice, some error or mistake might have occurred, giving some corrupt data or insights.

After getting the transaction rows, some things were combined on product level and number. If some mistake was already made in the database, a miss calculation and thus corrupt row was entered in the datasheet on the wrong place.

Due to the fact that the datasheet was of a very large proportion, only the investors the investors were selected that had one execution only account, with at least a known gender and age, used a private account and at least one of the three variables, education, general experience and work experience was known.

The comparison is done with a 90 percent winsorized mean dataset. This means that 10 percent of the data was adjusted. After all adjustments this is again working the dataset towards the desired output. But even after using the winsorized 90 percent mean, the predictive value was only 26.77%. Although the decision tree gives X, taking in mind all the assumptions and limitations, insights in the performance with its competitors. This could although change all the time, as the prices might change, implicating that the conclusions drawn in this research are only reliable on the moment of writing.

Although the model isn't as predictive as we wished it was, X could use it for marketing and targeting purpose. Although overall the predictive value was disappointing, the predictive value of some nodes and the distinction between some nodes was quite clear. X could check whether these factors are useful for marketing purpose.

The data used here is based on the year 2016, 205 and 2014 and is tested via the use of crossvalidation. One should remind that the year 2016 was the year in which Great Britain decided to leave the European Union and the year that Donald Trump was elected in the United States. Above that the increase threat of Islamic State was felt in Europe and the interest rates were nearly zero. All these events influence the dataset. Although the data is combined with the data of 2015 and 2014, resulting in balancing these events. More years would make the data more and more trustworthy and result in even more reliability.

7.3 Further research

The dataset provided many insights in the characteristics of investors. Although the forecasting model isn't as reliable as one hoped it would be, the research and dataset provided lots of fields that could be analysed in the future.

With the use of this dataset, one could investigate per age (group), what is found to be the most invested category per year on the stock exchange. One might investigate what persons per increase in wealth tend to invest in compared with those in other levels. Or investigate the preferred products for those with primary education and do this also for those on another education level. The same applies for the experience level of investors and the gender. This could then be combined with the marketing targets of X and thus help to specify their target group (or at least back up the ideas).

The dataset provides lots of chances to investigate the performance of X versus its competitors on one specific products category (for example on stocks). The outcome of these comparison can then also be used to see what prices changes would have done with the cost of these investors and thus give some insights in price changes and might thus help in getting the prices more accurate and suited to the investors within X.

For the decision tree and model developed itself, we could in hindsight have investigated the use of machine learning. "Machine learning is a discipline focused on two interrelated questions: How can one construct computer systems that automatically improve through experience? And what are the fundamental statistical-computational-information-theoretic laws that govern all learning systems, including computers, humans and organizations (Jordan and Mitchell, 2015)?" With the use of machine learning, the model could have automatically have updated or might have filled up during the year(s) and increased its predictive value. We suggest to read the paper of Jordan and Mitchell (2015) to gain more insights in machine learning.

Literature, figures and tables

Literature table

AFM (2013). "Report on leveraged products". Autoriteit Financiële Markten, Amsterdam.

Ameriks, J. and Zeldes S. (2004), "How do household portfolio shares vary with age?", working paper, Columbia University.

Anderson, A. (2013), "Trading and Under-Diversification," working paper, Institute for Financial Research, Stockholm.

Bali, T. G., Engle, R. F., & Murray, S. (2016). *Empirical asset pricing: the cross section of stock returns*. John Wiley & Sons.

Barber and Odean. (2011) The behaviour of individual investors

Barber, B.M. and T. Odean (2000), "Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors," Journal of Finance 55:773- 806.

Barber, B.M. and T. Odean (2001), "Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment,c Quarterly Journal of Economics 116:261-292.

Barber, B.M., Y. Lee, Y. Liu, and T. Odean (2007), "Is the Aggregate Investor Reluctant to Realize Losses? Evidence from Taiwan," European Financial Management 13:423- 447.

Barber, B.M., Y. Lee, Y. Liu, and T. Odean (2009), "Just How Much Do Individual Investors Lose by Trading? Review of Financial Studies 22:609-632.

Barber, B.M., Y. Lee, Y. Liu, and T. Odean (2011), "The Cross-Section of Speculator Skill: Evidence from day trading,"

Barberis, N. and R.H. Thaler (2003), "A Survey of Behavioral Finance," in G. Constantinides, M. Harris, R. Stultz eds., Handbook of the Economics of Finance. (North-Holland: Amsterdam), 1051-1119.

Calvet, L.E., J. Campbell, and P. Sodini (2009), "Fight or Flight? Portfolio Rebalancing by Individual Investors," Quarterly Journal of Economics 124:301-348.

Campbell, J.Y. (2006), "Household Finance," Journal of Finance 61:1553-1604.

Doorn, P.K. and Rhebergen, M.P. (1998), "Correlatie en Regressie; statistiek voor historici," Instituut voor Geschiedenis, Universiteit Leiden. Retrieved from: http://www.let.leidenuniv.nl/history/RES/stat/html/les10.html

Dorn, A.J., D. Dorn, and P. Sengmueller (2014), "Trading as Gambling," Working paper, University of Amsterdam Business School.

Dorn, D. and G. Huberman (2005), "Talk and Action: What Individual Investors Say and What They Do," Review of Finance 9:437-481.

Dorn, D. and P. Sengmueller (2009), "Trading as Entertainment," Management Science 55:591-603.

Feng, L., and M. Seasholes (2008), "Individual Investors and Gender Similarities in an Emerging Stock Market," Pacific-Basin Finance Journal 16:44-60.

Gao, X. and T. Lin (2010), "Do Behavioral Needs Influence the Trading Activity of Individual Investors? Evidence from Repeated Natural Experiments

Garson, G. D. (2012). Testing statistical assumptions. *Asheboro, NC: Statistical Associates Publishing*.

Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, *48*(2), 241-251.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 81-124.

Graham, J.R., C.R. Harvey, and H. Huang (2009), “Investor Competence, Trading Frequency, and Home Bias,” Management Science 55:1094-1106.

Grinblatt, M. and M. Keloharju (2000), “The Investment Behavior and Performance of Various Investor Types: A Study of Finland's Unique Data Set,” Journal of Financial Economics 55:43 67.

Grinblatt, M. and M. Keloharju (2001), “What Makes Investors Trade?” Journal of Finance 56:589-616.

Grinblatt, M., M. Keloharju, and J. Linnainmaa (2010), “IQ, Trading Behavior, and Performance,c Journal of Financial Economics, forthcoming.

Grinblatt, M., M. Keloharju, and J. Linnainmaa (2011), “IQ and Stock Market Participation," Journal of Finance, forthcoming.

IBM (2016). IBM SPSS Decision Trees 24. IBM Corporation. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/24.0/en/client/Manuals/IBM_SPSS_Decision_Trees.pdf.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

Keller, C., and M. Siegrist – “Money attitude Typology and Stock Investment”, Journal of Behavioral Finance, 7:2, (2006), pp. 88-96.

Korniotis, G.M. and A. Kumar (2009a), “Do Older Investors Make Better Investment Decisions?" Review of Economics and Statistics 93:244-265.

Korniotis, G.M. and A. Kumar (2009b), “Do Portfolio Distortions Reflect Superior Information or Psychological Biases?"

Lewellen, W.G., R.C. Lease, and G.G. Schlarbaum (1977), “Patterns of Investment Strategy and Behavior among Individual Investors,” 50:296-333.

Lim, V.K.G., and Teo, T.S.H. (1997), “Sex, Money and Financial Hardship: An Empirical Study of Attitudes Towards Money among Undergraduates in Singapore.” Journal of Economic Psychology, 18, 4, (1997), pp. 369–386.

Millward Brown (2014). “Retail Investor 2014, Ontwikkelingen op de markt van particuliere beleggers in Nederland 1997 – 2014.” Millward Brown.

Odean, T. (1999), “Do Investors Trade too Much?" American Economic Review 89:1279-1298.

Pallant, Julie. (2007). SPSS survival manual: a step by step guide to data analysis using SPSS. Maidenhead: Open University Press/McGraw-Hill.

Rezelman, E. (2015, November 10). Particuliere beleggers keren massaal terug naar de beurs. RTL Z Nieuws. Retrieved from http://www.rtlnieuws.nl/economie/home/particuliere-beleggers-keren-massaal-terug-naar-de-beurs.

Tukey, J. W. (1977). Exploratory data analysis.

Wood, R., and J.L. Zaichkowsky (2004), “Attitudes and Trading Behavior of Stock Market Investors: A Segmentation Approach.” Journal of Behavioral Finance, 5, (2004), pp. 170–179.

Figures

Tables

Appendix

Appendix A – Literature matrix

number	Gender	Age	Wealth	Diversi-fication / Stake size	Experience / knowledge	University /schooling /education	investment domestic equities	Gambling preference	Overconfidence and performance	IQ	Marital status
1	x		x	x	x	x			x		
2			x	x							
3	x		x	x							
4					x						
5			x	x	x						
6	x	x	x		x			x	x		x
7			x	x	x						
8				x			x				
9		x	x	x		x					
10		x	x								
11		x	x		x	x					
12	x	x	x	x	x	x					
13					x			x			
14	x	x	x		x	x		x			
15	x	x	x	x			x				
16			x	x				x			
17	x	x	x			x	x		x		
18							x				
19	x						x				
20		x	x		x					x	
21		x	x	x		x				x	
22	x	x	x			x		x			
23		x	x	x	x	x					
24		x		x		x	x				
25		x				x					x
26					x						
27									x		
28	x	x	x					x	x		
count	**10**	**15**	**19**	**13**	**12**	**11**	**6**	**6**	**5**	**2**	**2**

01. Anderson, A. (2013), "Trading and Under-Diversification," working paper, Institute for Financial Research, Stockholm.

02. Barber, B.M., Y. Lee, Y. Liu, and T. Odean (2009), "Just How Much Do Individual Investors Lose by Trading? Review of Financial Studies 22:609-632.

03. Barber, B.M., Y. Lee, Y. Liu, and T. Odean (2007), "Is the Aggregate Investor Reluctant to Realize Losses? Evidence from Taiwan," European Financial Management 13:423- 447.

04. Barber, B.M., Y. Lee, Y. Liu, and T. Odean (2011), "The Cross-Section of Speculator Skill: Evidence from day trading,"

05. Barber, B.M. and T. Odean (2000), "Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors," Journal of Finance 55:773- 806.

06. Barber, B.M. and T. Odean (2001), "Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment,c Quarterly Journal of Economics 116:261-292.

07. Barber and Odean. (2011) The behaviour of individual investors

08. Barberis, N. and R.H. Thaler (2003), "A Survey of Behavioral Finance," in G. Constantinides, M. Harris, R. Stultz eds., Handbook of the Economics of Finance. (North-Holland: Amsterdam), 1051-1119.

09. Calvet, L.E., J. Campbell, and P. Sodini (2009), “Fight or Flight? Portfolio Rebalancing by Individual Investors,” Quarterly Journal of Economics 124:301-348.

10. Ameriks, John, and Stephen Zeldes, 2004, How do household portfolio shares vary with age?, working paper, Columbia University

11. Campbell, J.Y. (2006), “Household Finance,” Journal of Finance 61:1553-1604.

12. Dorn, D. and G. Huberman (2005), “Talk and Action: What Individual Investors Say and What They Do,” Review of Finance 9:437-481.

13. Dorn, A.J., D. Dorn, and P. Sengmueller (2014), “Trading as Gambling,” Working paper, University of Amsterdam Business School.

14. Dorn, D. and P. Sengmueller (2009), “Trading as Entertainment,” Management Science 55:591-603.

15. Feng, L., and M. Seasholes (2008), “Individual Investors and Gender Similarities in an Emerging Stock Market,” Pacific-Basin Finance Journal 16:44-60.

16. Gao, X. and T. Lin (2010), “Do Behavioral Needs Influence the Trading Activity of Individual Investors? Evidence from Repeated Natural Experiments

17. Graham, J.R., C.R. Harvey, and H. Huang (2009), “Investor Competence, Trading Frequency, and Home Bias,” Management Science 55:1094-1106.

18. Grinblatt, M. and M. Keloharju (2000), “The Investment Behavior and Performance of Various Investor Types: A Study of Finland's Unique Data Set,” Journal of Financial Economics 55:43-67.

19. Grinblatt, M. and M. Keloharju (2001), “What Makes Investors Trade?” Journal of Finance 56:589-616.

20. Grinblatt, M., M. Keloharju, and J. Linnainmaa (2010), “IQ, Trading Behavior, and Performance,c Journal of Financial Economics, forthcoming.

21. Grinblatt, M., M. Keloharju, and J. Linnainmaa (2011), “IQ and Stock Market Participation," Journal of Finance, forthcoming.

22. Keller, C., and M. Siegrist – “Money attitude Typology and Stock Investment”, Journal of Behavioral Finance, 7:2, (2006), pp. 88-96.

23. Korniotis, G.M. and A. Kumar (2009a), “Do Older Investors Make Better Investment Decisions?" Review of Economics and Statistics 93:244-265.

24. Korniotis, G.M. and A. Kumar (2009b), “Do Portfolio Distortions Reflect Superior Information or Psychological Biases?"

25. Lewellen, W.G., R.C. Lease, and G.G. Schlarbaum (1977), “Patterns of Investment Strategy and Behavior among Individual Investors,” 50:296-333.

26. Lim, V.K.G., and T.S.H. Teo. “Sex, Money and Financial Hardship: An Empirical Study of Attitudes Towards Money among Undergraduates in Singapore.” Journal of Economic Psychology, 18, 4, (1997), pp. 369–386.

27. Odean, T. (1999), “Do Investors Trade too Much?" American Economic Review 89:1279-1298.

28. Wood, R., and J.L. Zaichkowsky. “Attitudes and Trading Behavior of Stock Market Investors: A Segmentation Approach.” Journal of Behavioral Finance, 5, (2004), pp. 170–179

Appendix B – Literature worked out, but out of scope

2.1.10 Gambling preference

Barber et al. (2009) found that people are driven by a gambling preference. In their sample of Taiwan, they found that when legal gambling was introduced in Taiwan, the turnover was reduced on the TSE by about one-fourth. Gao and Lin (2014) investigated Taiwan more in depth and found that when the jackpot exceeds 500 million Taiwan dollar, the trading volume decreases between 5,2% and 9,1% among stocks preferred by individual investors and between 6,8% and 8,6% among lottery-like stocks. The decline in individual buy volume is statistically indistinct from the decline in sell volume

Barber and Odean (2001) distinct two aspects of gambling, the risk seeking and entertainment. Risk seekers may account for under-diversification. Men and to a lesser extend women might also trade for entertainment, those who enjoy trading believe, overconfidently, that they have trading ability.

Dorn et all. (2015) found evidence from three different samples consistent with investors substituting between playing the lottery and gambling in financial markets. In the United States, increases in the jackpots of the multistate lotteries are associated with significant reductions in small trade participations in the stock market. Also California-based discount broker clients and German discount brokerage clients are less likely to trade during weeks of larger lottery prices. Dorn and Sengmueller (2009) investigated among 1000 German brokerage clients for whom both survey responses and actual trading records are available, investors who report enjoying investing or gambling turn over their portfolio at twice the rate of their peers. Meaning investors are trading to gamble and that they tend to trade more than those getting less satisfaction out of gambling. Keller and Siegrist (2006) found that only the risk-seekers group have a positive attitude towards gambling. They tend to tolerate financial risk well.

2.1.11 Overconfidence and performance

Anderson (2013) found that many investors trade too frequently and most of them perform below their self-selected benchmark portfolio. Their behaviour is puzzling and difficult to rationalize using normative financial theory. Some explanations have been offered, namely that investors may be overconfident and have an illusion of control.

Barber and Odean (2001) tested if overconfident investors trade excessively. They found in accordance with psychological research that men are more confident than women and thus that men trade more than women do. Their performance is worse than that of women investors. Graham et all. (2009) also found that male investors, and investors with larger portfolios or more education, are more likely to perceive themselves as competent than are female investors, and investors with smaller portfolios or less education.

Odean (2000) tested the hypothesis that investors trade excessively because they are overconfident. Overconfident investors may trade even with their expected gains through trading are not enough to offset trading costs. When trading costs are ignored, these investors actually lower their returns through trading. Wood and Zaichkowsky (2004) even maid a main segment of the confident traders. These investors have high levels of confidence and control. They are slightly older than the other groups, with 97% older than thirty, and their portfolio values are for 50% more than $100.000. 37% trade more than ten times per year. They tend to own the most stocks. As they trade more often, their investment horizon is of a shorter term.

2.1.12 IQ

Grinblatt et al. (2010) investigated whether IQ influences trading behaviour, performance, and transaction costs. The analysis combined equity return, trade, and limit order book data with two decades of scores from an intelligence (IQ) test administered to nearly every Finnish male of draft age. Grinblatt et al. (2010) found that high-IQ investors exhibit superior market timing, stock-picking skill, and trade execution. High-IQ investors bear no additional cost that offsets their stock-picking advantage. Further high-IQ investors' portfolio holdings outperform low-IQ investors' portfolios, especially when adjusted for differences in market timing.

Grinblatt et al. (2011) found that high-IQ investors are more likely to hold mutual funds and larger numbers of stocks and experience lower risk. The high correlation found between IQ and participation, which exists even among the 10% most affluent individuals, controls for wealth, income, age, and other demographic and occupational information. Grinblatt et al. (2011) found with supplemental data from siblings, studied with an instrumental variables approach and regressions that control for family effects, demonstrate that IQ's influence on participation extends to females and does not arise from omitted familial and nonfamilial variables.

2.1.13 Marital status

Barber and Odean (2001) found that the differences in portfolio turnover and net return performance to be larger between the accounts of single men and single women than between the accounts of married men and married women. Married couples tend to influence each other's investment decisions and thereby reduce the effects of gender differences in overconfidence.

Lewellen et al. (1977) found that "single investors, who are 65 or older, seek income more heavily than do persons whose families still contain at least one other member. This is, in so respects, counter to what we might anticipate at lower ages and is most likely a pension and/or social security phenomenon – that is, if there is a spouse who is either drawing federal benefits or has an additional private pension from previous employment, the relative current investment income need of the combined household is reduced.

Appendix H2 – Correlation matrix 99th percentile

	EXPERIENCE DUE TO WORK	EDUCATION	GENERAL EXPERIENCE	AGE	WEALTH	GENDER	PERCENTAGE COST
EXPERIENCE DUE TO WORK	1.000	0.999***	-0.099***	-0.016**	-0.054***	0.062***	-0.053***
EDUCATION	0.999***	1.000	-0.102***	-0.023***	-0.053***	0.059***	-0.056***
GENERAL EXPERIENCE	-0.099***	-0.102***	1.000	0.048***	0.071***	-0.005	0.000
AGE	-0.016**	-0.023***	0.048***	1.000	0.325***	0.096***	-0.083***
WEALTH	-0.054***	-0.053***	0.071***	0.325***	1.000	0.050***	-0.145***
GENDER	0.062***	0.059***	-0.005	0.096***	0.050***	1.000	-0.087***
PERCENTAGE COST	-0.053***	-0.056***	0.000	-0.083***	-0.145***	-0.087***	1.000