Predicting student enrollment

Logistic regression on attended marketing events

Master Thesis

Yente Hamers

Student-number: U1256596

ANR-number: 140390

Master Data Science: Business and Governance

Tilburg University

2016/2017

Supervisor: dr. G.A. Chrupała

External supervisor: A. Wanders

Second reader: dr. E.E. van der Vaart

Contact external supervisor:

Words: 15,931

Preface

This thesis is the last step in achieving my master's degree Data Science: Business and Government. In this section, I would like to express my gratitude to two people who helped me along the process of writing this master thesis.

First, I would like to thank my supervisor Grzegorz Chrupała, who took time for me, gave me useful feedback and was willing to answer all my questions. It was a pleasant learning process thanks to you.

I would like to express another big thanks to Alex Wanders, who made the dataset used in this research available. Alex helped me in the beginning of this entire process by brainstorming with me and putting forward some great ideas for my research. In addition, I could always count on a quick response on my questions. Thank you, Alex!

Summary

Nowadays there is a highly competitive market for higher education, which makes student recruitment an increasingly important topic for universities all over the world. It is important for universities to know how they can effectively attract new students using marketing efforts. Therefore, the goal of this research is to create a weighing mechanism in which it can be defined which marketing efforts play a role in deciding to study at Tilburg University and what role the different marketing efforts play. In addition, it is investigated which marketing efforts are the most important in predicting student enrollment and whether there are differences in the role of marketing efforts for potential bachelor and potential master students. This is, however, not the first research into the influence of various marketing efforts on student enrollment. There have been several researches that were often conducted through surveys that were only taken from confirmed students. In addition, these past researches only looked at which marketing efforts play a role in deciding to study at a university but not at what role the different marketing efforts play.

In this research, a dataset is used which was retrieved from the CRM system of the marketing department of Tilburg University. This dataset contains 39,619 confirmed and potential students and eleven different marketing efforts that are being used at Tilburg University. Different logistic regression models were trained, tuned and tested on different subgroups of this dataset. The models that yielded the highest model performance were used to retrieve the coefficients from. In addition, feature ablation analyses and bootstrap analyses were performed.

The results from the analyses show that the different marketing efforts can have a positive or negative relation with the enrollment in a degree at Tilburg University. In addition, it can be concluded that the on-campus events where face-to-face contact is possible are of more importance for the enrollment than online marketing efforts. In addition, it appears that university-specific marketing events do have a more positive relation with the enrollment than marketing events were more universities are present. In addition, the results of this research show that there are some differences in weights of marketing efforts for potential bachelor and potential master students.

Further research could investigate whether these findings are also applicable to the marketing efforts of other universities. Moreover, qualitative research could be done to look

deeper into why some marketing events have a negative relation and some events have positive relation with student enrollment.

Table of Contents

1. Introduction

This first chapter of the thesis provides a brief outline of the context of the project. Thereafter, a brief motivation for the project is given and the research questions are disclosed. Then, the practical and scientific relevance of the project is given and the further outline of the paper is briefly explained.

1.1. Context of project

There are fourteen different universities in the Netherlands, one of them is Tilburg University located in the Dutch city Tilburg. The foundations of the education at Tilburg University are quality of education, social orientation and a strong campus feel. Tilburg University currently has 12,399 students from 102 different nationalities. This higher education institute offers 21 different bachelor programs and 48 distinctive master programs in various disciplines. The different faculties, corresponding to the various disciplines, of Tilburg University are: Tilburg School of Economics and Management, Tilburg Law School, Tilburg School of Social and Behavioral Sciences, Tilburg School of Humanities and Tilburg School of Catholic Theology. Besides education, research is an important part of Tilburg University. The research of the different disciplines together contributes to 'Understanding society' and solving social problems (www.uvt.nl).

Every year students graduate and new students are recruited and welcomed to the campus of Tilburg University. The recruiting of new students is important for the survival of the university, after all, there is no need for education if there are no students. In addition, the number of students has a significant impact on the budget of the university. The recruitment of these new students is the responsibility of the marketing department of Tilburg University. This department uses a lot of different marketing efforts to inform and recruit new students. Examples of the marketing efforts are: provide brochures on request, organize information evenings, organize taster days, organize open days, answer online questions of potential students and send personalized e-mails.

To keep track of the potential students (potential students are the persons who have attended one of the marketing efforts online), the marketing department of Tilburg University uses a Customer Relationship Management system (CRM). This CRM system stores all available information about the students of Tilburg University and its potential students. For

example, when a potential student registers for an open day, this information is stored in the CRM system and this potential student becomes a lead. When this potential student also registers for an information evening, that information is added to this lead. In that way, the marketing department of Tilburg University can keep track of their potential students and their customer-journey.

1.2 Motivation

Student recruitment is an increasingly important topic for universities all over the world (Obermeit, 2012). Globalization of the economy, rapid technological changes and demographic diversity of the market are challenges higher education institutes need to face nowadays. These trends cause a highly competitive market for higher education (Maringe & Mourad, 2012). Universities worldwide struggle to position themselves in this highly competitive market and experience difficulties in recruiting new students (Hemsley-Brown & Oplatka, 2006; Maringe, 2006).

To stand out in this market with lots of competition, the needs of the students must be assessed. As competition among universities intensifies, a need for thorough understanding of the potential students becomes a vital step toward developing an efficient and effective marketing program (Coccari & Javalgi, 1995). Therefore, innovative ways and methods are needed to gather more information about potential students and which information sources they use (Maringe, 2006; Veloutsou, Paton & Lewis, 2005). Marketing efforts need to be more effective and efficient than ever before to reach the prospective students.

In addition, about 40% of the revenues of universities depends on the number of enrollments and graduates in the Netherlands (Meza, 2012). This is a significant part of the fixed income of universities. The number of students does not only directly affect the income of the university by tuition fees but also indirectly. In the Netherlands, the government has a fixed budget for higher education which is divided proportionately among the various higher education institutes. So, if a university has more students, it will not only gain more tuition fees but it will also gain more subsidies. To gain more subsidies, universities and colleges must not only try to attract more students, they need to attract more students than other institutes. This competition for the relatively most students, creates an even more competitive market for higher education (Meza, 2012).

1.3. Research questions

The data collected and stored by the CRM system are at the moment, despite their possible predictive value, not extensively used by the marketing department of Tilburg University. The marketing department does currently not know which marketing efforts are more important in the decision to start a degree at Tilburg University and which role the different marketing efforts play in this decision. At the moment, the weights of importance that are assigned to the different marketing efforts are self-invented based on common sense. However, this might not give a realistic view on what the weights and roles of the different marketing efforts really are. Therefore, the goal of this project is to create a weighing mechanism. In this weighing mechanism, it can be defined which marketing efforts play a role in deciding to study at Tilburg University and what role they play.

To achieve the goal of this research, the following research questions have been compiled:

Q1: *How can student enrollment be predicted based on attended marketing events?*

Q2: *How are the different marketing efforts affecting the decision of potential students to enroll in a degree at Tilburg University?*

Q3: *Which marketing efforts are most important and which marketing efforts the least in predicting student enrollment?*

Q4: *What are the differences in weights of the marketing efforts in the decision to enroll between potential bachelor and potential master students?*

All the above formulated research questions can be answered by creating a logistic regression model. With this model, predictions could be made and the weights of the different marketing efforts could be assessed. By answering the second research question, insights are gathered about the size and direction of the weights of the different marketing efforts. The third research question will look at the influence on model performance of the different marketing events singly to investigate which marketing events do have the most predictive power for this prediction task.

1.4. Relevance

1.4.1. Practical relevance

Since the marketing department of Tilburg University has limited financial resources for recruiting new students, this project might be of great practical benefit. When the marketing department of Tilburg University has a more accurate view on the role of the marketing efforts in the decision to study at the university, the limited financial resources can be used in a smarter and more efficient way (Maringe, 2006). If one marketing effort appears to be very important in the decision to enroll in a degree at Tilburg University, the marketers should pay more attention to this marketing effort. On the other hand, if a marketing effort appears to play no, or a small, role at the decision to study at Tilburg University, it might be wise to cut the budget on that marketing effort.

Moreover, if the recruiting of new students becomes more efficient and effective and it results in more enrollments, the budget of Tilburg University rises. As mentioned in section 1.2, the budget of the university is directly and indirectly influenced by the number of enrollments. This might lead to a higher budget for the marketing department but also for research and publications, which might affect the prestige and reputation of the university. This might in turn lead to even more enrollments.

In addition, this research might not be only of practical benefit for the marketing department of Tilburg University but also for other marketing departments of higher education institutes. Other marketing departments of higher education institutes might carry out a similar research based on this study or might generalize the findings of this research to their own practice.

1.4.2. Scientific relevance

This research might, in addition to the practical value, also be of value from a scientific point of view. There has been only a few studies into, and academic attention to, branding and recruitment in higher education (Bock, Poole & Joseph, 2014). Moreover, the research that has been done up until now, regarding how students chose their higher education institute, is more focused on characteristics of universities like reputation and job perspectives (Kim & Gasman, 2011; Pampaloni, 2010; Maringe, 2006; Briggs, 2006). There has not yet been a lot of research, to the best of my knowledge, into the consequences of the marketing efforts itself in combination with universities. And if it concerns a study into the consequences of the

marketing efforts itself, it often mainly concerns interpersonal effects like the effect of parental influence. Moreover, most of the research focuses on countries or particular universities but not on Tilburg University or the Netherlands.

This project may for these reasons be an enrichment of the known academic literature because it views student recruitment from a different angle than what has already been done in this field of research.

1.5. Outline of this document

In this document, the following structure is adhered to. Chapter 1 gives an introduction on the subject of the thesis and its practical and scientific relevance. Moreover, the research questions are defined in the first chapter. Chapter 2 provides the theoretical framework, containing an overview of relevant literature on the effect of marketing efforts in student recruitment. Chapter 3 describes the dataset used, as well as the processes of data collection, data cleaning and data merging. In chapter 4 the approach, mathematical formulas and algorithm which is used in this document are described. In chapter 5 the experiments are described. Chapter 6 presents the results. Chapter 7 provides a discussion and describes the implications of this research. In chapter 8, the conclusions and suggestions for further research are given.

2. Background

This chapter gives a brief overview of what research has already been done in the field of higher education marketing and student recruitment. First of all, the area of research is briefly explained. The second section describes why a degree at a higher education institute can be categorized as a high-involvement service. Thereafter, the process of information gathering is described and a distribution of information sources is explained. Then different marketing efforts which create offline and online communication channels are discussed. After that, four ranking-studies are explained. And at last, the current research is discussed.

2.1. Area of research

This current research, as defined in chapter 1, can be defined in the field of higher education marketing. Interest in the field of higher education marketing is growing rapidly due to changes in the market (Maringe & Mourad, 2012). Due to these changes universities worldwide struggle to position themselves in this highly competitive market and experience difficulties in recruiting new students (Hemsley-Brown & Oplatka, 2006; Maringe, 2006).

There are many different subjects of research in the higher education marketing field. A lot of academic literature is about the university attributes that attract students and about personal characteristics that influence the choice of university (Pampaloni, 2010; Rowan-Kenyon, Bell & Perna, 2008; Bartl, as cited in Obermeit, 2012; Whitehead, Raffan & Deaney, 2006; Kim & Gasman, 2011; Briggs, 2006). This research however focuses on the effect of marketing efforts on the decision-making process of prospective students. In the following sections, the scope of the current research is delimited and defined.

2.2. High-involvement service

A degree at a higher education institute can be categorized as a high-involvement service, since the potential student has an elevated level of concern and attention for the service. This prominent level of concern arises because wrong purchase decisions can have financial implications and can force the student to deal with a poor service for an extended period (Laurent & Kapferer, 1985). For example, the decision to attend university is one that the student is highly involved with, as this would impact greatly on their future life and prospects (Drummond, 2004). In addition, the level of involvement reflects on the information gathering

process (Drummond, 2004); the higher the product or service involvement, the more time a consumer spends searching for product information to make the right decision regarding the product or service (Liang, 2012; Beatty & Smith, 1987). In other words, customers have a more engaged purchase decision when it concerns a high-involvement product or service (Gu, Park & Konana, 2012) and engage in extensive information search (Beatty & Smith, 1987).

2.3. Information gathering

Since a degree at Tilburg University is a high involvement service, information retrieval is very important for the potential students (Beatty & Smith, 1987). The information search of prospective students covers the process undertaken to acquire information and identify viable solutions for their problem. This information gathering can be internal or external (Blackwell, Miniard & Engel, 2006). Internal information search is based on retrieving existing information in memory or knowledge from previous related experiences. External search is information beyond one's personal sources. External search occurs when internal search is not perceived as sufficient for a confident decision and whenever the benefits of the additional information prevail over its costs (Schmitt & Spreng, 1996).

Understanding the external information gathering process of the prospective students has a major importance and reflection on the promotional strategy of an institution (Blackwell, Miniard & Engel, 2001). It is important for marketers to discover the kind of information the prospective student wishes to know. Different studies indicate that potential students search among other things for information about: the financial aspects regarding a degree at that institution (Pampaloni, 2010; Rowan-Kenyon et al., 2008), the courses that are offered by the institution (Bartl, as cited in Obermeit, 2012; Whitehead, et al., 2006), distance of the institution from home (Kim & Gasman, 2011; Briggs, 2006), location of the institution (Maringe, 2006; Kim & Gasman, 2011; Pampaloni, 2010), atmosphere of the city where the institution is located (Bartl, as cited in Obermeit, 2012) and the reputation of the institution (Kim & Gasman, 2011; Pampaloni, 2010; Maringe, 2006).

However, not only the kind of information the prospective students wish to know is important for the marketers. The marketers also need to know from what sources the prospective students collect their information (Blackwell, et al., 2001): the information sources.

2.4. Different information sources

Most potential students try to gain external information from different information sources at least one year prior to enrollment (Cabrera & La Nasa, 2000) but also when they are not planning to buy the product immediately (ongoing search) (Bloch, Sherrell & Ridgway, 1996). According to Brown, Varley and Pal (2009) there are three major categories of information sources. These three categories of information sources differ in by whom they are created and controlled. The three categories of information sources are: non-controllable sources of information, partly controllable sources of information and controllable sources of information (Brown, et al., 2009).

Non-controllable sources of information are provided by external sources, without influence and control of an internal party of the higher education institute. These information sources cannot be influenced by the university, even if the information provided is false or negative (Brown, et al., 2009). Maringe (2006) showed that reviews about the faculty and program quality of an institution, which are non-controllable sources of information, are of influence on the enrollment-decision of potential students. Another study showed that also the advice of relatives, friends and teachers regarding which institution to choose is of influence on the enrollment-decision of a student (Kim & Gasman, 2011; Rowan-Kenyon, et al., 2008).

Partly controllable sources of information are combinations of sources. This means that most of the time both an external and an internal party have influence on the information source (Brown, et al., 2009). Examples of these kind of sources are regional and national newspaper publications, popular magazine articles and social networks.

Controllable sources of information originate from and are controlled by the educational institutions themselves. Examples of these kind of information sources are: informative brochures, booklets, open days, taster days and website-content (Brown, et al., 2009).

All three kinds of information sources do have a considerable influence on the decision to enroll in a university. However, only the controllable sources of information are of importance for the university marketers and recruiters because they can influence or create these information sources directly by different marketing efforts. In addition, several studies proved that prospective students primarily rely on information sources developed by the

university. According to these studies, the partly controllable and non-controllable sources of information are of less importance (Veloutsou, Lewis & Paton, 2004; Briggs & Wilson, 2007).

2.5. Marketing efforts

Once the information sources have been identified, the marketer can begin promotional effort to build communication channels to provide prospective students with more information (Blackwell, et al., 2001). It is recommended that marketers attempt to motivate students to collect more information about their institutions (Al-Fattal, 2010).

There are different marketing efforts that are related to controllable information sources and thus communication channels. These marketing efforts are of significant importance for university marketers because retrieving information is a very big part of the decision-process of a potential student to enroll in a university. Moreover, the choice for different marketing efforts is very important for its success. Marketers must prevent information overload because too much information can be a problem for customers, even when it involves a high-involvement product or service (Schwartz, 2000).

The different marketing efforts that are related to controllable information sources, and can be used to create communication channels, can be divided in online and offline marketing efforts.

2.5.1. Offline marketing efforts

Although the internet is the main source of information seeking nowadays (Kim & Gasman, 2011; Pampaloni, 2010; Rowan-Kenyon, et al., 2008; de Jager & du Plooy, 2010), face-to-face contact appears to be more influential than high-tech sources of information. Moreover, information retrieved in a face-to-face conversation is more accessible and has a profound influence on judgements (Herr, Kardes & Kim, 1991). Moony and Robben (1997) suggest that by increasing a direct experience with the product (the institution), the opportunity to process more information will increase and lead to a more favorable evaluation.

Hosting open days, taster days and open lectures are examples of offline marketing efforts related to controllable information sources. Physically visiting a school weighs heavily in the decision-making process to enroll in a university (Pampaloni, 2010; Moogan, Baron & Harris, 1999; Omboi & Mutali, 2011). Open days, taster days and open lectures allow students

to ‘see the school with their own eyes’ and allow them to answer questions such as ‘Am I welcome?’ and ‘Would I fit in?’ (Moogan et al., 1999). These events are opportunities for the institutes to communicate a positive image to the prospective students (Fischbach, 2006).

In addition, being able to ask questions to school representatives in person is very important in the decision-process of prospective students (Johnston, 2010; Moogan, et al., 1999). For many prospective students, having contact with a school representative provides the information they feel they need to have to decide about going to study at that school (Pampaloni, 2010). The questions of potential students can be answered at open days, taster days and open lectures but also at organized education fairs.

According to research of Simões and Soares (2010), education level has influence on which marketing efforts are of most effect on the decision to enroll. According to them, prospective students with the highest education level tend to refer more to interpersonal sources.

2.5.2. Online marketing efforts

Different studies show that the main source of information seeking by potential students nowadays is on the internet (Kim & Gasman, 2011; Pampaloni, 2010; Rowan-Kenyon, et al., 2008; de Jager & du Plooy, 2010). Therefore, the marketing efforts of higher education institutions needs to shift from print to the web in order to reach the prospective students. Although it is hard to retrieve emotional information like ‘Do I feel welcome at this university?’ online for potential students, online marketing efforts can have influence on the students decision-making process.

One of the possible online marketing efforts related to controllable information sources is to provide online brochures. Brochures and other informative publications from and about universities are very important information sources for prospective students (Briggs, 2006; Moogan & Baron, 2003; Veloutsou et al., 2005; Briggs & Wilson, 2007). Prospective students look to prospectuses and brochures for detailed information (Connor, Pearson, Pollard, Tyers & Willison, 2001). However, these brochures and publications do only have a positive effect on the decision-making process if the information is complete, detailed and easy to understand (Briggs, 2006; Moogan & Baron, 2003).

Another possible online marketing effort for creating communication channels is providing the potential students an informative website. Web pages can easily be monitored

and maintained for image, information and accuracy (Abrahamson, 2000). In research of Simões and Soares (2010) 81% of the prospective students indicated that the university website was the most used information source. Students in this research seemed to resort less to university brochures and leaflets.

Other examples of online marketing efforts related to controllable information are to provide prospective students the possibility to ask an online question to a school representative and to provide prospective students videos about the university. In that way, prospective students might retrieve information that they otherwise, without a visit to the campus, not would have retrieved.

2.6. Rankings of information sources

Different studies investigated the importance of the different information sources for prospective students (Johnston, 2010; Pampaloni, 2010; de Jager & du Plooy, 2010; Simões & Soares, 2010). These four studies investigated, among other things, the importance of different information sources in the decision-making of prospective students. These researches are much alike in their method and their subject but are conducted with prospective students from different universities and in different countries. Research of de Jager and du Plooy (2010) focused on South African students, research of Simões and Soares (2010) focused on Portuguese students and Pampaloni (2010) and Johnston (2010) investigated American students. These four different studies have similarities and differences in the information sources they have investigated.

To investigate the different information sources, all four studies made use of questionnaires with open-ended questions and questions on a Likert-scale. In these questionnaires, it was explicitly asked which information sources were of most influence and which information sources were most used. The number of respondents variated per study: Jager and du Plooy (2010) had 410, Simões and Soares (2010) had 1,641 respondents, Pampaloni (2010) had 249 respondents and Johnston (2010) had 475 respondents. Based on the completed questionnaires, all four of these studies created a ranking of importance for the information sources. In Table 1 the four different rankings are showed, derived from the four different studies.

Although the four studies are much alike, the results are quite different. Table 1 shows for instance that in studies of de Jager and du Plooy (2010), Simões & Soares (2010) and Pampaloni (2010) 'University website' appears to be a very important information source (in top 3 of ranking). However, in research of Johnston (2010), 'University website' is ranked on the 10th place and thus, appears to be not important at all.

Another striking difference in the results of the four studies is about the importance of 'Campus visits & open days'. As can be seen from Table 1, 'Campus visits & open days' has a top 3 ranking in the studies of de Jager and du Plooy (2010), Simões and Soares (2010) and Johnston (2010). However, in the results of Pampaloni (2010) 'Campus visits & open days' appears to be not important and receives only a 9th ranking.

Three of the four studies incorporated in Table 1, also investigated the importance of 'University publications (leaflets & brochures)'. For this information source, the different studies showed approximately the same result. In research of Jager and du Plooy (2010), Simões and Soares (2010) and Pampaloni (2010) 'University publications (leaflets & brochures)' got ranked on the 4th and 5th place.

The differences in ranking could be explained by the fact that the United States of America, South Africa and Portugal are geographically quite far apart. In addition, the differences in rankings could be explained by the cultural differences that might be present between the different countries.

Rank	De Jager & du Plooy (2010)	Simões & Soares (2010)	Pampaloni (2010)	Johnston (2010)
1	University website	University website	Campus visits & open days	University staff
2	Campus visits & open days	Other students (alumni)	High school teachers	Parents
3	High school teachers	High school teachers	University website	Campus visits & open days
4	University publications (leaflets & brochures)	University official guides	Interview at the school	Friends
5	Parents	University publications (leaflets & brochures)	University publications (leaflets & brochures)	Family members
6	Word-of-mouth	Promotional visits to secondary school	High school counsellors	High school teachers or counselors
7	Events on campus	High school counsellors	Media/publicity	University sport events
8	Advertisement in magazines/newspapers	Media/publicity	Recruiters	Paper mail
9	Other students (alumni)	Campus visits & open days	University official guides	Social network sites
10	Advertisements on the radio	Other	College fair	University official guides
11	Advertisements on TV		Family/friends	University website
12	Other		Clergy	E-mail from university

Table 1: *Rankings from 4 different studies concerning information source importance.*

Although these four different studies provide valuable insights, they have some implications (Johnston, 2010; Pampaloni, 2010; de Jager & du Plooy, 2010; Simões & Soares, 2010). First, all four of the studies have been conducted by means of a survey. In these surveys, the respondents were explicitly asked about the use of different information sources and their importance. In that way, it is assumed that the students have a very good memory and self-awareness of this complex decision. In addition, in all the four studies confirmed students were asked about their decision to enroll in a university, not real prospective students. Therefore, reminders of the complex decision might even be worse because it has been a

while. Another limitation of the four studies is the generalizability. All four studies were conducted on one out of different universities in different countries and at different faculties. There might be differences between choosing for university A and B but there also might be difference between choosing a management degree at university A and university B.

Lastly, the four different studies do have some similarities in their rankings but also some striking differences. These differences might be explained by the geographical distance between the countries of investigation and/or by the cultural differences. However, the studies of Johnston (2010) and the research of Pampaloni (2010) both were conducted in the United States and do also show some major differences. So, the existing literature does not have a clear answer to the question which information sources are the most important for prospective students.

2.7. Current research

This current research will focus on the marketing efforts related to controllable information sources and their predictive value in the decision whether to enroll for a degree at Tilburg University. The goal of this research is to determine which marketing efforts play a role in the decision to study at Tilburg University and what that role is.

As discussed in section 2.6, different studies have already been investigating the importance of different information sources. Although these different studies (Johnston, 2010; Pampaloni, 2010; de Jager & du Plooy, 2010; Simões & Soares, 2010) are comparable in their design, they do show different results. These differences in results might be due to the fact that each of the four studies investigated a different university in another country. This kind of study has not yet been conducted for a university in the Netherlands and therefore this study on Tilburg University might be an enrichment of the existing literature.

This current study has some similarities with the four studies discussed in section 2.6. This study will also investigate the importance of different information sources and make a kind of ranking. However, there are also relevant differences between this current study and the studies of Johnston (2010), Pampaloni (2010), de Jager and du Plooy (2010) and Simões and Soares (2010).

First of all, this study investigates different kind of information sources than the information sources investigated in the four discussed studies. This current research does, unlike the other researches, not focus on interpersonal information sources. The much

discussed ‘parental influence’ (Cabrera & La Nasa, 2000; Kim & Gasman, 2011; Rowan-Kenyon et al., 2008) is, for example, not included in this research.

Secondly, in this research a distinction is made between potential bachelor and master students. In the discussed ranking-studies, distinctions have been made but not a distinction based on education level. The bachelor-master distinction is incorporated in this research because there is a suspicion that bachelor and master students have different information gathering processes.

Furthermore, this research will not only create rankings of importance of the different information sources, it will also attach weights to them. In that way, a more detailed view of the importance of the different information sources is provided and what their role is in the decision-making process.

Finally, in contrast with the discussed ranking-studies, this research will not be based on questionnaires. In this work, the information will be learned from data using a logistic regression model. There will be a logistic regression on the attended marketing events to be able to predict whether a potential student is likely to enroll in a degree at Tilburg University or not. By using a logistic regression model there is no influence of memory and self-awareness of the prospective students, it will only be based on real used information sources (attended marketing events). In addition, the data this research is based upon is not limited to a small sample and contains as well confirmed as non-confirmed students.

3. Data

In the first section of this chapter, the data collection is described. Thereafter, the process of merging the two collected datasets is described. Then, the data cleaning process is explained. The last section of this chapter provides a detailed description of the final dataset with tables and figures.

3.1. Data collection

Two datasets have been used for this research which were extracted from the CRM system of the marketing department of Tilburg University. One of these datasets contained the potential students of the university that had attended one of their marketing-events (leads dataset) and the other dataset contained the confirmed students of the university (student dataset). The confirmed students were also in the leads dataset; however, the information of these confirmed students was missing.

Both datasets did contain about the same features, however, the student dataset did also contain the name of the program the student finally registered for, the start- and predicted end date and the nationality of the student.

3.2. Data merging

The confirmed students in the leads dataset had missing values for almost all features. Therefore, the two datasets had to be merged so that the information of the student dataset would also be in the leads dataset. The merging and cleaning of the two datasets was performed in RStudio.

To merge the leads and the student dataset, the corresponding features were named alike. Unfortunately, there was no corresponding ID-number that could connect the rows in the leads dataset to the rows in the student dataset. Therefore, the two datasets were joined on e-mail address which resulted in a dataset with some duplicates. Table 2 shows an example of the dataset after merging the two datasets. In addition, the duplicates that retained after merging were filtered out were possible, sometimes manually to retain as much information as possible. However, despite careful inspection, it might be that there are still duplicates in the final dataset (e.g. when using two totally different e-mail addresses). Unfortunately, this is insurmountable.

E-mail	Confirmed	Level	Brochure	Taster.Day.Ba	Open.Day.Ba
Anonymous1@gmail.com	1	Bachelor	NA	NA	NA
Anonymous1@gmail.com	1	Bachelor	0	1	0
Anonymous2@gmail.com	0	Master	1	2	0
Anonymous3@gmail.com	0	Bachelor	0	0	1

Table 2: *Small example of the dataset after merging two datasets.*

3.3. Data cleaning

After merging the datasets into one, the dataset was cleaned: all string values were made lower case, numbers and odd characters were removed, string factors were turned into numerical factors and some rows with completely wrong values were deleted. Subsequently, the missing values were made NA values. The rows with NA values for level of education (Bachelor/Master) were removed from the dataset because these would not be informative for this research. These rows could not be divided into the master and bachelor subsets. In addition, the column 'Student.Status', which contains the nationality of the students, was removed from the dataset. This feature was removed because the values of the potential students were different from the values of the confirmed students. For example, the confirmed students often had the labels 'National' and 'International' while the not confirmed students had labels as 'Potential international' and 'Potential national'. These 'Potential national' and 'Potential international' labels were not completely trustable because these labels were derived from the attended marketing events. For instance, when a Dutch student visits an international education fair it is likely to be labeled as 'Probably International'. Thereafter the first names, last names and email addresses were removed from the dataset to make it anonymous.

In addition, a new feature was created: distance from place of residence to the campus. The distance to the campus variable was created using the longitude, latitude and city features. To gather as much information as possible, some longitudes and latitudes were entered manually per city. With a function created in RStudio, the distances in kilometers were computed and added as a feature.

However, despite careful cleaning and maintaining as much information as possible, there are a lot of missing values in the dataset for the 'Distance to campus' feature. Therefore, this 'Distance to campus' feature was removed from the final dataset, to make the algorithm operate faster and more effectively (Kotsiantis, Zaharakis & Pintelas, 2007). However, a

smaller dataset was created in which only the rows without NA values for 'Distance to campus' were maintained. This smaller dataset, of 14,739 rows, could later possibly be used for some separate analyzes to see if distance is of any influence on the enrollment of potential students.

At last a new column was created based on the sum of 11 count features. An example of one of these count features is the number of attended open bachelor days. This column was created to make sure that only the rows which had attended at least one of the marketing efforts would remain. This research focuses namely on predicting student enrollment based on attended events. Moreover, only the confirmed students who did attend zero events would be in the dataset, potential students namely only become a lead when they attend one of the events. For these reasons, students who did not attend one of these touchpoints were deleted because they would not be of value for this research.

No transformations were performed because the count features were already in the same format and there already was a limited number of features. However, adding feature interactions were considered and tested during the analyses.

3.4. Dataset description

The final dataset contains 39,167 rows and 11 count features. In Table 3 and Table 4 some descriptive statistics are displayed. Table 3 shows that there are more non-confirmed students in the dataset than confirmed students. Table 3 also shows that there is a reasonably equal distribution between Master and Bachelor students. In Table 4 the minimum value, maximum value and average is displayed per feature. The average in this fourth table means how often a single potential student attended that event on average. As can be noticed from Table 4, the most attended marketing event is the bachelor open day. The least attended marketing events, also displayed in Table 4, are the international visit and the open lecture for potential bachelor students.

Education level	Master	Bachelor	Research Master	PhD	Pre-Bachelor	Pre-Master	Total
Not confirmed students	15,817	11,338	929	43	9	343	28,479
Confirmed students	3,638	7,485	14	0	0	3	11,140
Total	19,455	18,823	943	43	9	346	39,619

Table 3: *Descriptive statistics confirmed and non-confirmed students and education level.*

Feature	Minimum value	Maximum value	Average	Occurrence
Open.Day.Ba.Count	0	6	0.32	12,693
Online.Questions.Count	0	13	0.25	9,948
Brochure.Count	0	3	0.15	5,731
Open.Day.Ma.Count	0	3	0.13	5,156
International.Fair.Count	0	1	0.11	4,279
Taster.Day.Ba.Count	0	5	0.07	2,880
Webinar.Count	0	3	0.07	2,574
Education.Fair.Count	0	1	0.01	214
Taster.Day.Ma.Count	0	2	0.01	340
International.Visit.Count	0	1	0.00	28
Open.Lecture.Ba.Count	0	2	0.00	30

Table 4: *Minimum value, maximum value and average per feature.*

The dataset contains five education level related marketing events: an open lecture for prospective bachelor students, a bachelor taster day, a master taster day, a bachelor open day and a master open day. Figure 1 shows the distribution of attendees on these five education level related marketing events and the non-education level related marketing events. As can be seen from Figure 1, the master-related marketing events are well attended by both prospective master and bachelor students. This equal distribution, however, does not apply when it comes to bachelor-related events. As can be seen from Figure 1, the master taster days are even more often visited by potential bachelor students than prospective master students. In addition, it can be seen in Figure 1 that the online videos and online lectures are more attended by master students than by bachelor students.

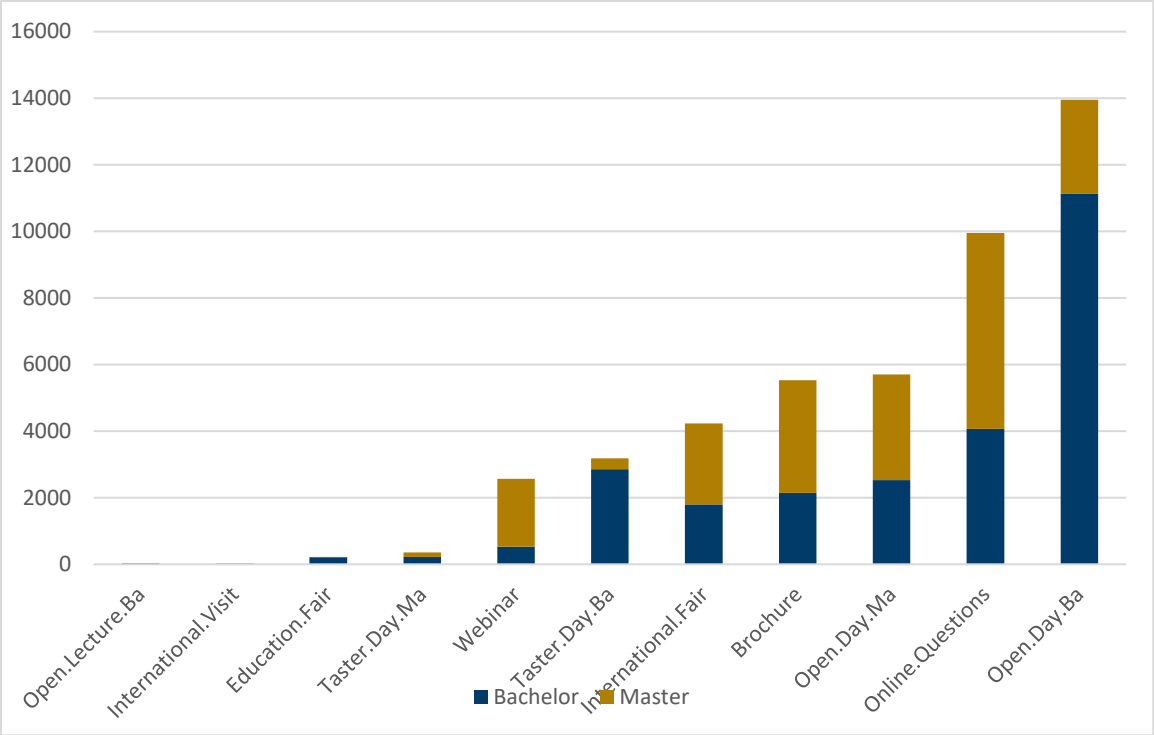

Figure 1: *Total number of attendees per education level on marketing events.*

Figure 2 shows the total number of attended marketing events per person. As can be seen from Figure 2, the most potential students in the dataset did attend only one marketing event. A minority of the potential and confirmed students attended two or more marketing events.

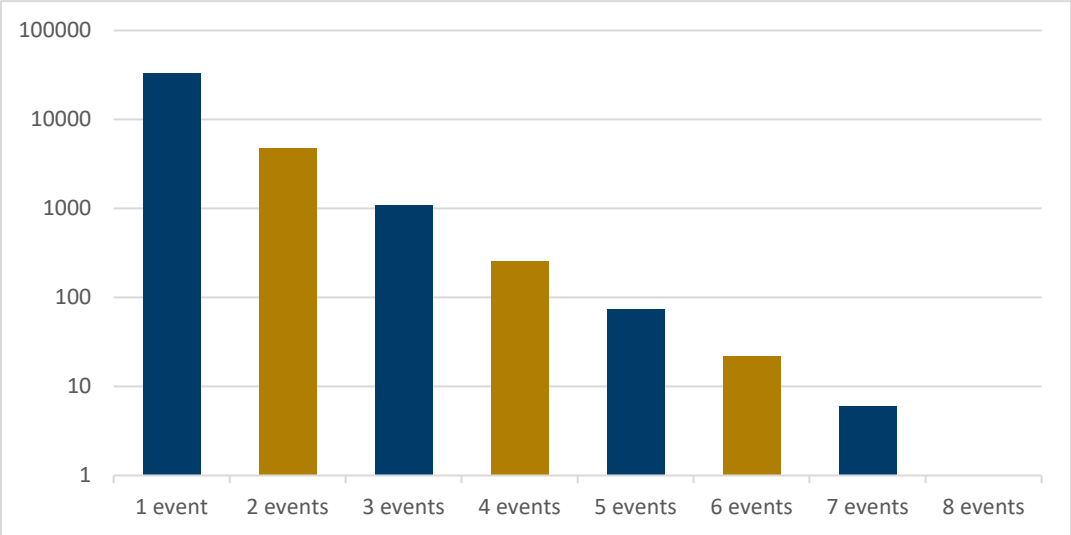

Figure 2: *Histogram with total attended marketing events*.

In addition to the final dataset, a subset with the 'Distance from campus' feature was created. Based on the longitude and latitude data in this subset, Figure 3 has been created in R. Figure

3 shows the worldwide distribution of prospective students and confirmed students of Tilburg University. As mentioned before, the longitude and latitude data were not available for all the students and potential students in the dataset, so only a smaller part of the data is visible in the map of Figure 3. However, it can still be noticed from Figure 3 that there is not only national interest in a degree at Tilburg University but also international interest from all parts of the world.

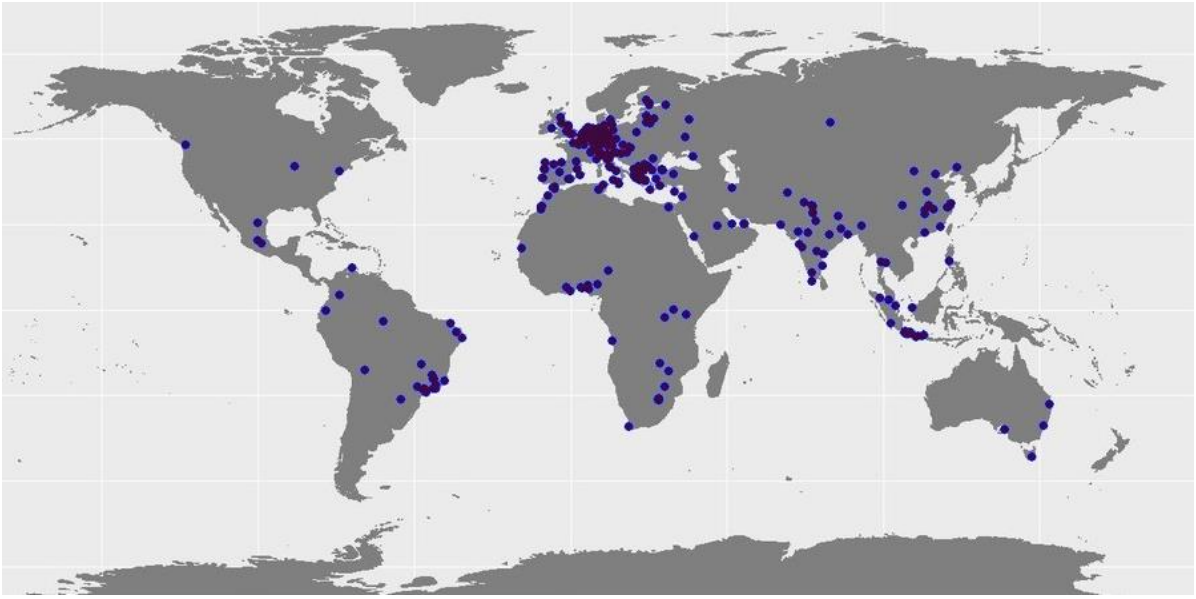

Figure 3: *Distribution of students and prospective students worldwide.*

However, it can also be noticed from Figure 3 that most of the confirmed students and prospective students are from Europe since there is an accumulation of dots. Because Figure 3 shows an accumulation of dots in Europe and Tilburg University is located in the Netherlands, the distribution of confirmed and potential students over the Netherlands is shown in Figure 4. From Figure 4 can be seen that most of the Dutch confirmed and prospective students of Tilburg University come from the south of the Netherlands, mainly from the provinces Limburg and Noord-Brabant.

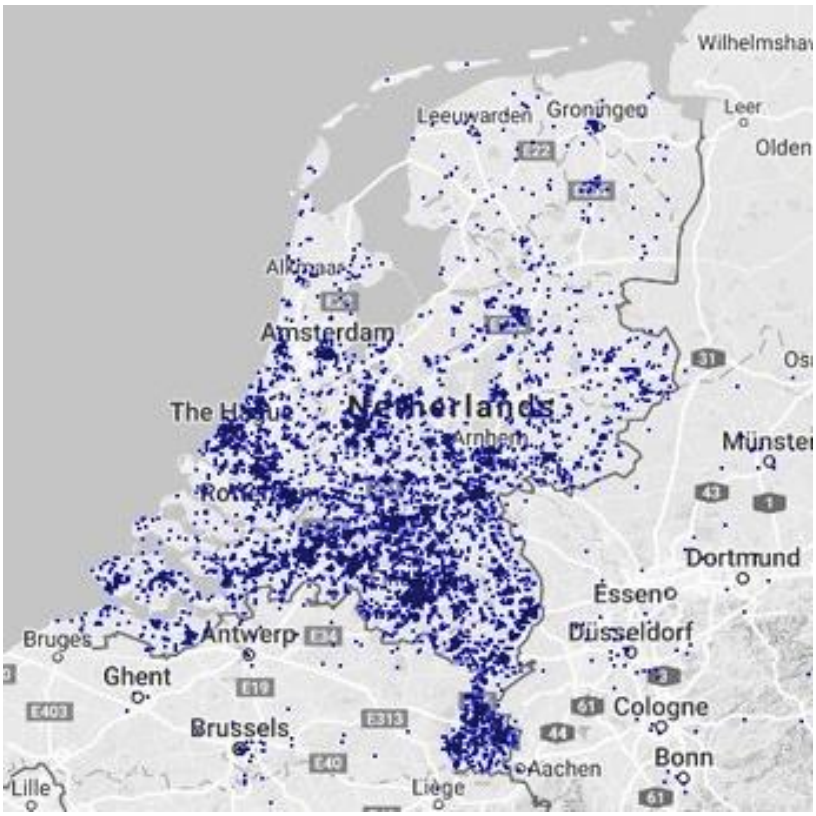

Figure 4: *Distribution of students and prospective students over the Netherlands.*

Table 5 also shows the distribution of the created 'Distance to campus' feature. From Table 5 can be noticed that some confirmed students and prospective students live very close to the campus of Tilburg University (0.16km: Tilburg, the Netherlands) and some confirmed students and prospective students live very far from the campus of Tilburg University (17,068.01km: Launceston, Australia). The average and standard deviation for 'Distance to campus' are both quite large, which might be due to presence of international students in the dataset.

Feature	Minimum value	Maximum value	Average	Standard Deviation
Distance to campus	0.16	17,068.01	272.68	1,255.40

Table 5: *Descriptive statistics about the distance to the campus of students and prospective students in kilometers.*

4. Method

In this chapter, the experimental procedure is discussed in more detail. First the variables of this study are briefly discussed. Thereafter, the experimental setup is described. In this experimental setup section, the experimental task, used algorithm and baseline are discussed in more detail. Thereafter the implementation of the algorithm is briefly explained. Then the relevant parameters of the algorithm are discussed and at last the evaluation criteria are explained. In the section of the evaluation criteria, the evaluation scheme of this project and the different error metrics used are addressed.

4.1. Variables

In the Appendix A: description of features, the distinctive features of this research are displayed. In the following two sections, a brief description is given.

4.1.1. Dependent variable

The dependent variable in this research is the 'Converted' feature. This dependent variable is binary and can have the value 0, non-confirmed student or 1, confirmed student. The 'Converted' feature measures whether a person has subscribed on StudieLink for a degree at Tilburg University or not. In this research, a person who subscribed to a degree at Tilburg University via StudieLink is considered to be a confirmed student. StudieLink is the Dutch national database for admission to higher education (www.uvt.nl).

Another option to measure what the influence of the different independent variables is, could be looking at the 'My.SAS' feature in the dataset. This feature measures whether a student has requested admission at the university. However, it was chosen to use the 'Converted' feature because in that case the university also accepted the student. So, the 'StudieLink' feature has been chosen as dependent variable because the prospective students that have subscribed to StudieLink are further in the process of enrolling for a degree at Tilburg university than the prospective students that have only subscribed to 'My.SAS'.

4.1.2. Independent variables

The dataset contains 11 different independent variables. These 11 independent variables are count features that are automatically created by the CRM system. These count features keep

track of how often a particular lead attends a specific marketing event, for instance an open day. These automatically created count features will be considered true and accurate in this research because no defects were found when comparing a small sample of the automatic counts to the manual counts. The manual counts were based on the dataset containing all registrations and applications for marketing events. In that dataset, every application or registration for a marketing event, is a separate row. This research aimed to look at what role these different independent variables have in predicting the dependent variable.

Three of the independent variables can be categorized as online information sources, as described in section 2.5.2. These online marketing events are: providing an online brochure, providing the possibility to ask questions online and providing online lectures and online videos about the university. How often a brochure has been requested is measured by the 'Brochure.Count' feature, how often an online question has been asked is measured by the 'Online.Questions.Count' feature and how often an online video about the university or an online lecture has been watched is measured by the 'Webinar.Count' feature.

The other eight independent variables can be categorized as offline information sources. How often a bachelor taster day has been visited is measured by 'Taster.Day.Ba.Count' and how often a master taster day has been visited is measured by 'Taster.Day.Ma.Count'. How often a prospective student visited a bachelor open lecture is measured by 'Open.Lecture.Ba.Count'. How often a bachelor open day has been visited is measured by the features 'Open.Day.Ba.Count'. 'Open.Day.Ma.Count' measures how often a master open day have been visited by a potential student. The features 'International.Fair.Count' and 'Education.Fair.Count' measure how often an (international) fair has been visited. Finally, 'International.Visit.Count' measures how often an registered individual international visit has been made.

4.2. Experimental setup

4.2.1. Experimental task

This research concerns a classification task. The goal of a classification task is to use a model or algorithm to predict the class of previously unseen examples or instances. This research involves binary classification: will the unseen example (unseen lead in the CRM system) become a student or not?

The learning of an experimental task can be supervised or unsupervised. In this research, the learning in the experimental task is supervised. Supervised learning means that

the examples are given with known labels, while with unsupervised learning the instances are unlabeled (Kotsiantis, et al., 2007).

4.2.2. Algorithm

In this research, the logistic regression classifier has been used. The logistic regression model has been chosen for multiple reasons. First of all, the logistic regression classifier is, like all classifiers, capable of making a binary classification which provides an answer to the first research question of this study: 'How can student enrollment be predicted based on attended marketing events?'. In addition, within a logistic regression model it is possible to extract the coefficients that are used to fit and build the model. The contribution of these coefficients can be interpreted: coefficient sizes determine their relative importance for the classification result (Dreiseitl & Ohno-Machado, 2002). Accessing and interpreting these coefficients might provide answers to the second, third and fourth research question as provided in chapter 1. Moreover, the logistic regression classifier outputs on the probabilities of classes. Therefore, a logistic regression model might make it possible to divide prospective students into various categories based on their probability of becoming a student.

As mentioned before, logistic regression is a supervised learning algorithm, which outputs on the probabilities of labels. Basically, the logistic regression classifier tries to find the best fitting logistic function for the data and then implements a threshold of 0.5 to divide the data into the two binary classes. Given a set of training examples, the logistic regression classifier learns a linear scoring function like in linear regression:

$$z = \boldsymbol{w} \cdot \boldsymbol{x} + b$$

where z is the score of the linear model, x is the feature value vector, w is the feature weight vector and b is the intercept. This linear function is used in logistic regression to predict logit(P):

$$logit(P_{pred}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$$

where P_{pred} is the predicted probability score, varying between 0 and 1. The heart of the logistic regression classifier is the inverse logit function, also known as the sigmoid function. This function is used to map the logit back to probability:

$$logit^{-1}(z) = \frac{1}{1 + \mathrm{e}^{-\mathrm{z}}}$$

where e is the numerical constant Euler's number and z is an input we plug into the function. This sigmoid function is then used in the final prediction function of a logistic regression classifier:

$$P_{pred} = logit^{-1}(z)$$

where P_{pred} is the predicted probability score and z an input we plug into the function.

The log-loss function, also known as cross-entropy, quantifies the mistakes of the logistic regression classifier. The logistic regression classifier tries to minimize the log-loss function to find the model which gives the maximum probability to the training targets. This log-loss function can be denoted as follows:

$$\ell_{log}(z) = -ylog\left(P_{pred}\right) - (1-y)\,log\left(1-P_{pred}\right)$$

Where $P_{pred} = logit^{-1}(z)$

where P_{pred} is the predicted probability score and y is the dependent variable which can be 0 or 1.

The model complexity of a logistic regression model is quite low, especially when no or only a few interaction terms and feature transformations are used. Therefore, overfitting is often not an issue in logistic regression (Dreiseitl & Ohno-Machado, 2002).

4.2.3. Baseline

The majority classifier was used on the dataset as baseline to predict whether a lead would become a student or not. The majority classifier takes the majority label in the training set and predicts that majority label for all new examples. This classifier was chosen as the baseline because the dataset used in this research is quite unbalanced, as previously seen in Table 3. When the majority classifier is used on an unbalanced dataset, the baseline accuracy will be quite high. However, the final algorithm should perform better than just assigning the majority class to every new instance, because otherwise it does not learn at all (Provost, 2000).

4.3. Implementation

Although the data merging and data cleaning were performed in R, the implementation of the logistic regression model and the baseline were carried out in Python. The code was written in a Jupyter notebook version 4.3.1 with Python version 3.

While cleaning and merging data in R, different packages were used: dplyr, ggmap and rworldmap. In addition, two different packages were used in the python code: Scikit-learn and

numpy. Scikit-learn offers simple and efficient tools for data mining and data analysis (www.scikit-learn.org). From the Scikit-learn module the train_test_split, LogisticRegression, PolynomialFeatures, GridSearchCV, StandardScaler and metrics were used. The numpy package was mainly used to read and load the data in the correct format into the Jupyter notebook.

4.4. Parameters

The LogisticRegression model of Scikit-learn used for this project has around 14 different parameters which can be tuned. Some of these parameters are relevant for this project and will be briefly explained in the next paragraphs.

The first relevant parameter of this model is the regularization parameter C. This C parameter controls the inverse of the regularization strength. The default value of C is 1 and C must always be a positive float. The smaller the value of C, the stronger the regularization.

Another relevant parameter of the logistic regression model is class_weight. This parameter might be of importance for this project because of the unbalanced dataset. The class_weight parameter can be set to 'none', 'balanced' and 'auto'. The 'balanced' mode of this parameter weighs classes proportionally to the inverse of their frequency. If class_weight is set to 'none' or it is not given at all, the classes are supposed to have weight one. Tuning this parameter might result in a better handling of the unbalanced dataset.

4.5. Bootstrap and feature ablation analysis

In this research, there has been made use of bootstrap analyses and feature ablation analyses. In the bootstrap analyses, the data have been resampled over training, test and validation set with replacement 50 times. After each resampling, the tuned logistic regression model was used to make predictions and the coefficients of the model are stored. Thereafter, the 50 sets of coefficients are analyzed. In that way, it is possible to determine whether the coefficients of the models are stable and reliable. A bootstrap analysis can approximate the sampling distribution.

To investigate the importance of the different marketing events for the enrollment of potential students, feature ablation analyses were performed. In a feature ablation analysis, the same model is performed multiple times, with every time another feature left out. By

comparing the model performance with all features with the model performance when one feature is left out, the importance of that left out feature can be determined.

4.6. Evaluation criteria

4.6.1. Evaluation scheme

One of the most important requirements for a machine learning project is the evaluation scheme. In this research, an experimental setup with three different sets was used: a training set, a validation set and a test set. Often the dataset is split in the way that there is two-third for training and what is left is evenly divided among test and validation set (Kotsiantis, et al., 2007). In this research, the training set consists of 60% of all the data, the validation set consists of 20% of all the data and the test set also consists of 20% of all the data. On the training set the model with different settings of parameters is trained, which is checked by the performance on the validation set. The final model is the model that performed best on the validation set. The test set is only used in the very last stage of the research to test the final model created with the training and validation set.

4.6.2. Error measure

The performance of the created model is assessed by examining the agreement between the target predictions and the actual targets. The confusion table in Table 6 can be used to calculate the upcoming four evaluation metrics.

	Predicted confirmed-student	**Predicted non-confirmed student**
Actual confirmed-student	True positives (TP)	False negatives (FN)
Actual non-confirmed-student	False positives (FP)	True negatives (TN)

Table 6: *Confusion Matrix.*

One of the most used error metrics is accuracy (Kotsiantis, et al., 2007). Accuracy assesses the overall performance of the created models. The accuracy is the percentage of correct predictions divided by the total number of predictions. In other words, the accuracy assesses the overall effectiveness of the algorithm (Sokolova, Japkowicz & Szpakowicz, 2006). The formula for calculating the accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP is the number of True Positives (correctly identified as potential student), FP is the number of False Positives (incorrectly identified as potential student), TN is the number of True Negatives (correctly identified as not potential student) and FN is the number of False Negatives (incorrectly identifies as not potential student).

5. Results

In this chapter, the results of the different analyses are discussed. First the performed majority classifier, GridSearch and the choice for the final model is discussed. Thereafter the results of the analyses on all education levels, potential bachelor students separately and potential master students separately are discussed. At last, the results of the additional feature 'Distance to campus' are reviewed.

5.1. Parameter tuning and model selection

For this research three different models were created: one for all education levels together (also including for instance potential pre-bachelor students), one for only bachelor students and one for only master students. First of all, to determine the baseline, the majority classifier was tested on the validation set. Table 7 shows the results of these analyses. As can be seen from Table 7, the accuracy scores of the baseline are already quite high. This can be explained by the unbalanced nature of the dataset. Table 7 shows that the dataset with only potential master students performs best on the majority classifier which means that 85% of the data are not confirmed students and only 15% are confirmed students.

	Accuracy
All education levels	0.74
Bachelor	0.61
Master	0.85

Table 7: *Accuracy on validation set with majority classifier.*

After determining the baseline for the three different datasets, the parameters of the logistic regression models had to be tuned. To optimize these three different models (all education levels, bachelor only and master only) on the validation data, GridSearch was used on the parameters C and class_weight as discussed in section 4.4. For all three models, two different sets of features were put into the GridSearch (one without feature interactions and one with feature interactions) to find out which one would be the best fitting model.

Table 8 shows the results of the performed GridSearch over the different models. When comparing Table 7 and Table 8, it can be concluded that all the logistic regression models shown in Table 8 yield higher accuracies than the baseline in Table 7. What also can be noticed from Table 8, is that adding feature interactions does, remarkably, not significantly increase the accuracy of the models. This might be explained by the fact that most of the

potential students and confirmed students only attended one of the eleven marketing events, as can be seen in Figure 2 in chapter 3. Because the feature interactions did not increase the accuracy and to keep the model less complex (Dreiseitl & Ohno-Machado, 2002), the models without interaction terms were selected as the final models.

		Accuracy	C-value	Class_weight
All education levels	**No interactions**	0.83	10	None
	2-degree interactions	0.84	1	None
Bachelor	**No interactions**	0.76	10	Balanced
	2-degree interactions	0.76	0.1	None
Master	**No interactions**	0.90	1	None
	2-degree interactions	0.89	0.1	None

Table 8: *Results of GridSearch for three models.*

In the next three sections, the results of these models will be discussed. These three separate sections will each consist of a description of the coefficients, a bootstrap analysis of the coefficients, a feature ablation analysis on the marketing events and a description of the model performance on the test set.

5.2. All education levels

The model for all education levels was trained on all available data, so not only potential bachelor and master students but also for instance potential pre-master students. To be able to answer the research questions, the coefficients of the logistic regression model were extracted. Table 9 shows the coefficients of the different marketing events based on the data with potential students from all education levels. As can be seen from Table 9, there are positive and negative coefficients. For example, the more international fairs a potential student has been visiting, the smaller the chance of an actual enrollment. However, the more frequent a master taster day has been visited by a potential student, the higher the chance that this potential student enrolls in a degree at Tilburg University.

	All education levels
International.Fair.Count	-3.30
Webinar.Count	-1.27
Open.Lecture.Ba.Count	-1.12
Brochure.Count	-1.02
Education.Fair.Count	-0.87
International.Visit.Count	0.26
Online.Questions.Count	0.49
Taster.Day.Ba.Count	0.72
Open.Day.Ba.Count	1.46
Open.Day.Ma.Count	2.80
Taster.Day.Ma.Count	8.17

Table 9: *Coefficients of features from the logistic regression model.*

To make sure whether these coefficients are reliable or not, a bootstrap analysis was performed. The outcome of this bootstrap analysis is shown in Figure 5. As can be seen from Figure 5, there are no big fluctuations in coefficients. In addition, there are no coefficients that switch from positive to negative when different training data is used to build the model on.

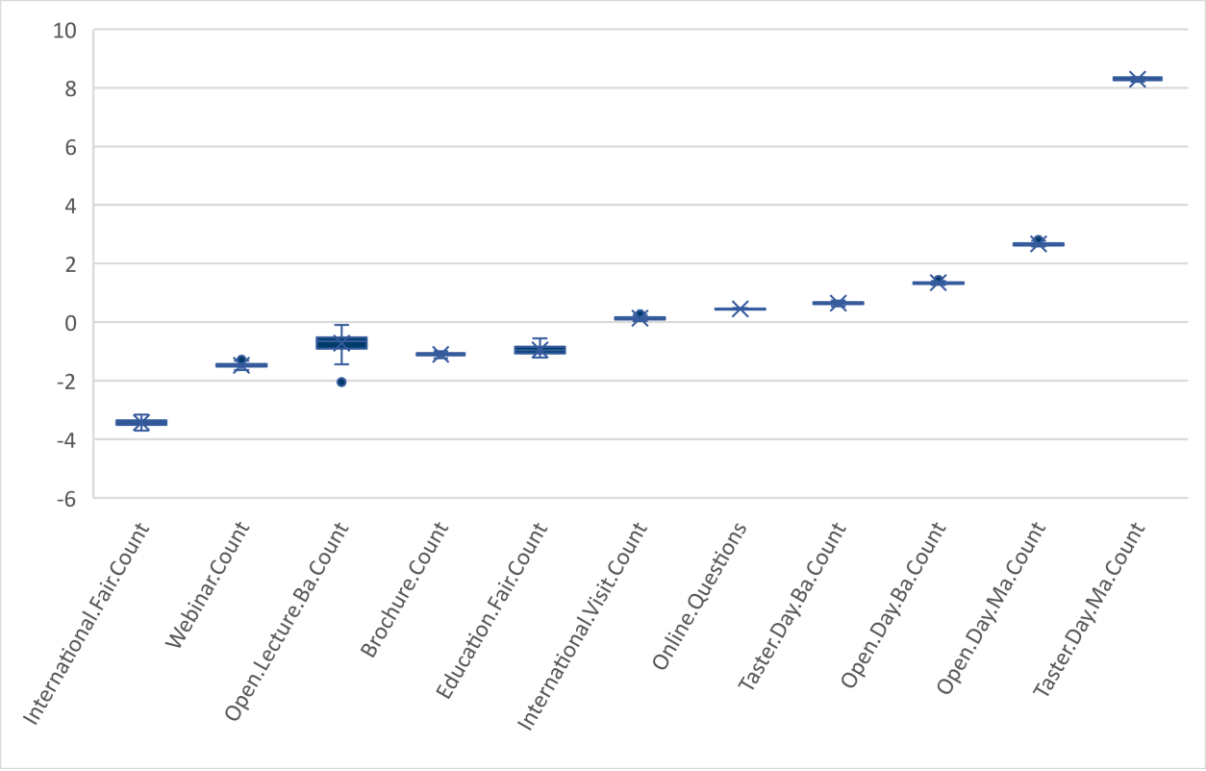

Figure 5: *Results of bootstrap analysis on coefficients of all education levels.*

The coefficients of Table 9 make it possible to create a ranking. Table 10 shows the ranking, in which number 1 has the relatively biggest coefficient and number 11 the smallest. It appears

from Table 10 that the master taster day has the biggest coefficient and the international visit the smallest.

Ranking	Marketing event	Weight
1	Taster.Day.Ma.Count	8.17
2	International.Fair.Count	-3.30
3	Open.Day.Ma.Count	2.80
4	Open.Day.Ba.Count	1.46
5	Webinar.Count	-1.27
6	Open.Lecture.Ba.Count	-1.12
7	Brochure.Count	-1.02
8	Education.Fair.Count	-0.87
9	Taster.Day.Ba.Count	0.72
10	Online.Questions.Count	0.49
11	International.Visit.Count	0.26

Table 10: *Coefficients of features from the logistic regression model.*

Thereafter, a feature ablation analysis was performed to investigate the importance of the different marketing events. The results of this feature ablation analysis are shown in Table 11. As can be seen in Table 11, there are some features that are not important when predicting student enrollment: asking questions online, visiting an international fair, making an international visit, visiting an open lecture for prospective bachelor students and watching online lectures or videos. However, it appears that the brochure, the education fair, the bachelor open day and the master and bachelor taster days are of little importance. According to Table 11, the master open day has the most predictive value for the predictive task.

	All education levels
Open.Day.Ma.Count	-0.06
Brochure.Count	-0.01
Education.Fair.Count	-0.01
Open.Day.Ba.Count	-0.01
Taster.Day.Ba.Count	-0.01
Taster.Day.Ma.Count	-0.01
Online.Questions.Count	0
International.Fair.Count	0
International.Visit.Count	0
Open.Lecture.Ba.Count	0
Webinar.Count	0
Accuracy all features	0.83

Table 11: *Difference in accuracy of model when the feature in the left column is left out.*

Finally, the accuracy of the logistic regression model and majority classifier were measured on the test set. The results of these two analyses can be found in Table 12. Table 12 shows that the logistic regression model has a higher accuracy than the baseline for both the validation and test set. In addition, it is noteworthy that the logistic regression model accuracy of the validation set and the accuracy for the test set are equal. However, it should be noted that although the logistic regression model has a higher accuracy than the baseline, it is only a very small improvement of 8%.

		Accuracy
Validation set	**Majority classifier**	0.74
	Logistic Regression	0.83
Test set	**Majority classifier**	0.75
	Logistic Regression	0.83

Table 12: *Accuracies of validation and test set for all education levels.*

5.3. Bachelor students

The model for only bachelor students was trained on data containing only the potential bachelor students. To be able to retrieve information about the differences between prospective bachelor and prospective master students, the analyses were also carried out separately for these two education levels. Table 13 shows the coefficients of the different marketing events for only prospective bachelor students. These coefficients were derived performing the tuned model as described in section 5.1. From Table 13 can be noticed that there are five negative coefficients and six positive coefficients. It appears that visiting an international fair is negatively related to enrolling for a degree at Tilburg University. This also applies to requesting a brochure, watching an online video or lecture and visiting an education fair. On the other hand, visiting a taster day or open day for prospective bachelor students are quite positively related to enrolling in a degree at Tilburg University. However, it is very remarkable that visiting a taster day and open day for prospective master students are even more positively related to enrolling into a bachelor degree at Tilburg University.

	Bachelor
International.Fair.Count	-3.23
Brochure.Count	-1.30
Education.Fair.Count	-1.03
Open.Lecture.Ba.Count	-0.93
Webinar.Count	-0.62
International.Visit.Count	0.13
Online.Questions.Count	0.58
Taster.Day.Ba.Count	0.60
Open.Day.Ba.Count	0.92
Open.Day.Ma.Count	5.41
Taster.Day.Ma.Count	7.27

Table 13: *Coefficients of features from the logistic regression model.*

Also for these bachelor coefficients, as displayed in Table 13, a bootstrap analysis has been done. The results of this bootstrap analysis can be seen in Figure 6. Figure 6 shows that there are no big fluctuations in the coefficients. According to Figure 6, the coefficients extracted from the logistic regression model for prospective bachelor students, are quite stable.

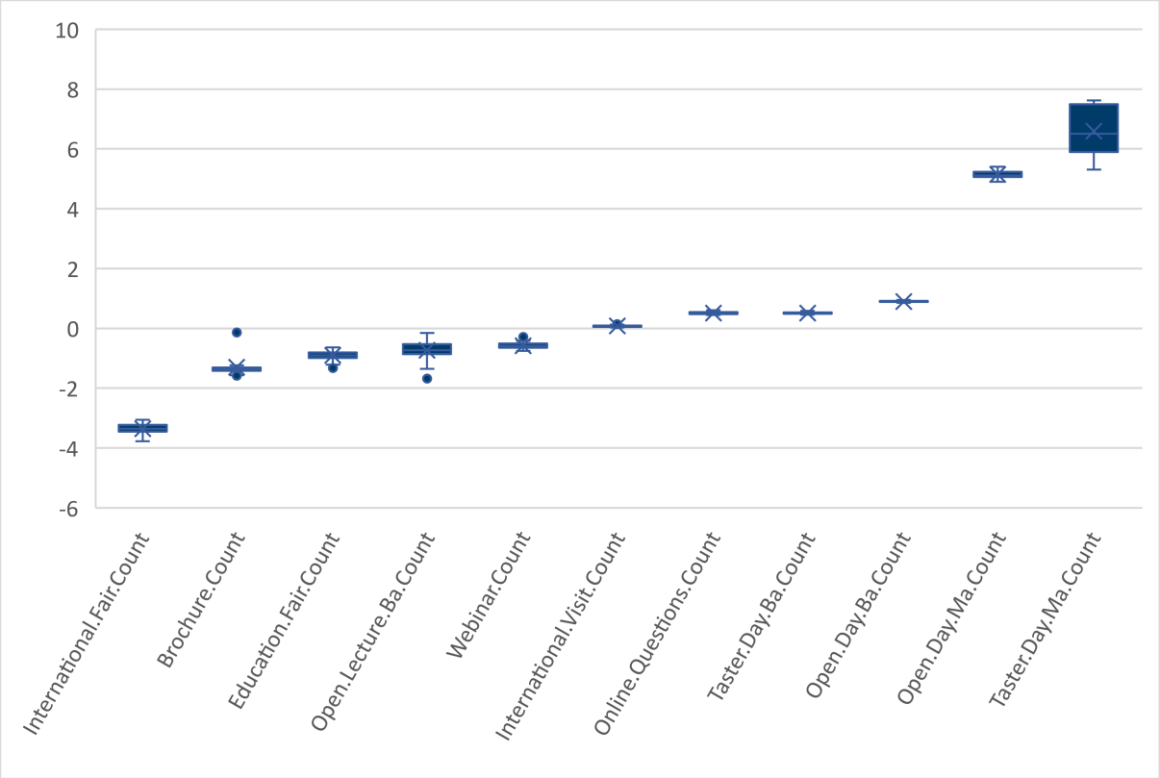

Figure 6: *Results of bootstrap analysis on coefficients of prospective bachelor students.*

In addition, a feature ablation analysis was performed on the bachelor student data. The results of this ablation analysis are shown in Table 14. From Table 14 can be noticed that the master open day is the most important feature in predicting whether a prospective bachelor student enrolls in a bachelor degree or not. The taster day and open day for prospective bachelor students, the master taster day, the brochure and the international fair are also of little importance to predict enrollment of prospective bachelor students. Furthermore, Table 14 shows that the six other marketing events separately do not have any predictive value for whether a prospective bachelor student enrolls in a bachelor degree at Tilburg University.

	Bachelor
Open.Day.Ma.Count	-0.18
Online.Questions.Count	-0.04
Taster.Day.Ba.Count	-0.04
Open.Day.Ba.Count	-0.02
Taster.Day.Ma.Count	-0.01
Brochure.Count	0
Education.Fair.Count	0
International.Fair.Count	0
International.Visit.Count	0
Open.Lecture.Ba.Count	0
Webinar.Count	0
Accuracy all features	0.76

Table 14: *Difference in accuracy of model when the feature in the left column is left out.*

Finally, the tuned logistic regression model and the majority classifier were performed on the test set. The results of these analyses can be found in Table 15. Table 15 shows that the baseline accuracy was the same for both sets. Both for the validation set and the test set the logistic regression model outperforms the majority classifier, even though this is only a small difference. In addition, it can be seen from Table 15 that the logistic regression model even performs 1% better on the test set than on the validation set.

		Accuracy
Validation set	**Majority classifier**	0.61
	Logistic Regression	0.76
Test set	**Majority classifier**	0.61
	Logistic Regression	0.77

Table 15: *Accuracies of validation and test set for only bachelor students.*

5.4. Master students

To be able to investigate the differences in weights of the different marketing events between master and bachelor students, also the master students separately were investigated. Table 16 shows the coefficients of the different marketing events that were extracted from the logistic regression model created for only the master student data. As can be seen in Table 16, potential students who visit an international fair or watch an online video are less likely to enroll as a master student at Tilburg University. Also, students who tend to request a brochure and/or visit a bachelor taster day are a little more likely to enroll in a degree at Tilburg University than students who do not attend one of these events. Table 16 also shows that visiting a bachelor or master open day and visiting a master taster day do have a considerable positive relation with the enrollment of prospective master students.

	Master
International.Fair.Count	-4.20
Webinar.Count	-2.06
Brochure.Count	-0.71
Taster.Day.Ba.Count	-0.27
Open.Lecture.Ba.Count	0.00
International.Visit.Count	0.15
Online.Questions.Count	0.31
Education.Fair.Count	1.48
Open.Day.Ma.Count	2.16
Open.Day.Ba.Count	2.69
Taster.Day.Ma.Count	5.66

Table 16: *Coefficients of features from the logistic regression model.*

To make sure that the coefficients as presented in Table 16 are reliable and stable, a bootstrap analysis has been performed. As can be seen in the boxplot of Figure 7, there are no large fluctuations in the coefficients. However, it can be noticed that the 'Taster.Day.Ba.Count' feature does fluctuate more than the other coefficients. This coefficient might be less reliable because it can have both a positive and a negative value.

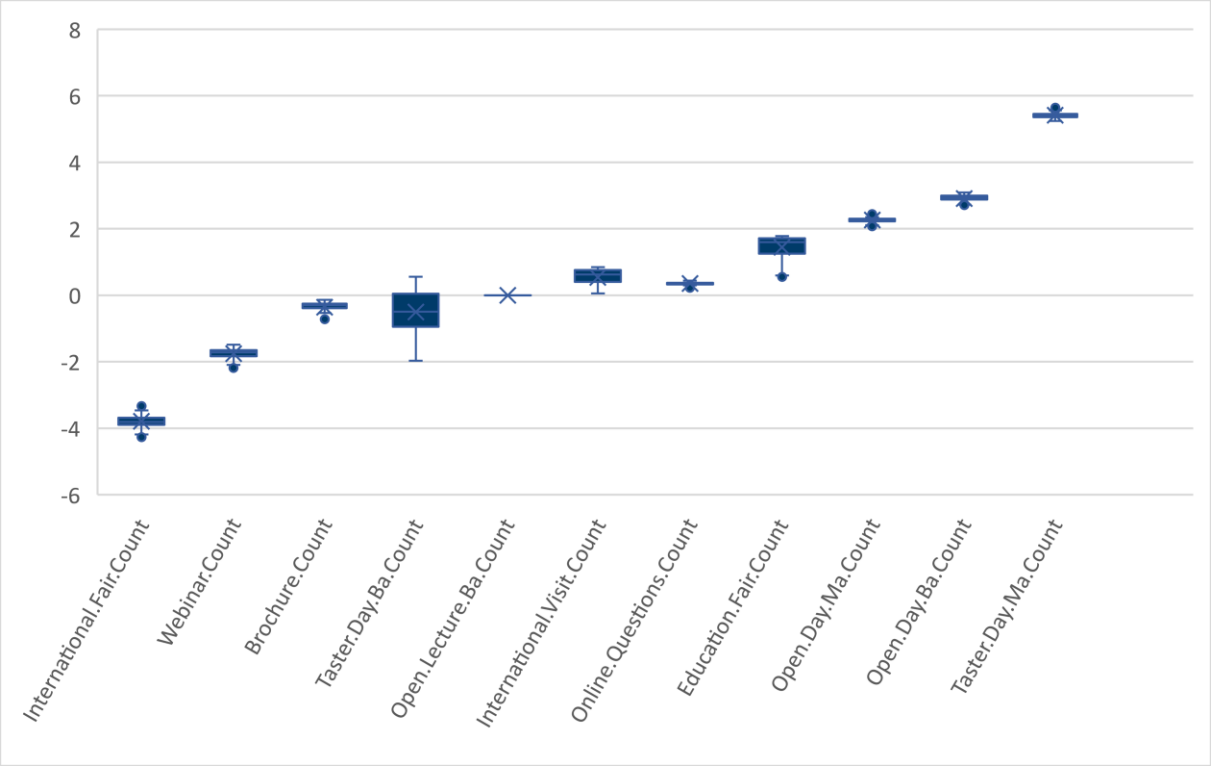

Figure 7: *Results of bootstrap analysis on coefficients of prospective master students.*

In addition, a feature ablation analysis was performed for the prospective master student coefficients. The results of this ablation analysis are shown in Table 17. According to Table 17, the bachelor open day is the most important marketing event for predicting the enrollment of master students. Making an international visit, visiting an open lecture for prospective bachelor students and visiting a bachelor taster day appear to have no predictive value for the enrollment of prospective master students. All the other marketing events do have a small predictive value for the prediction task.

	Master
Open.Day.Ba.Count	-0.03
International.Fair.Count	-0.02
Brochure.Count	-0.01
Online.Questions.Count	-0.01
Education.Fair.Count	-0.01
Open.Day.Ma.Count	-0.01
Taster.Day.Ma.Count	-0.01
Webinar.Count	-0.01
International.Visit.Count	0
Open.Lecture.Ba.Count	0
Taster.Day.Ba.Count	0
Accuracy all features	0.90

Table 17: *Difference in accuracy of model when the feature in the left column is left out.*

At last, the created logistic model and the majority classifier were performed on the test set. The results of these analyses can be found in Table 18. As can be seen in Table 18, the logistic regression model does perform 1% less accurate on the test set than on the validation set. In addition, the accuracy of logistic regression model on the test set scores 3% higher than the majority classifier.

		Accuracy
Validation set	**Majority classifier**	0.85
	Logistic Regression	0.90
Test set	**Majority classifier**	0.86
	Logistic Regression	0.89

Table 18: *Accuracies of validation and test set for only master students.*

5.5. Distance from campus

To test whether the distance from residence of the prospective students to the Tilburg University campus has influence on the enrollment of prospective students, the subset with only the examples with available information about their residence was analyzed. This used subset contained the 11 count features plus the 'Distance from campus' feature.

A performed GridSearch indicated that the logistic regression model with the default settings was the best fitting model for this subset of data. Table 19 shows that this logistic regression model outperforms the majority classifier for both the validation and test set but the difference is very small.

		Accuracy
Validation set	**Majority classifier**	0.78
	Logistic Regression	0.81
Test set	**Majority classifier**	0.78
	Logistic Regression	0.82

Table 19: *Accuracies of validation and test set for distance from campus subset.*

To investigate whether the 'Distance from campus' feature had any predictive value, a feature ablation analysis was performed. In this feature ablation analysis, the 'Distance from campus' feature was removed to see whether the performance of the model is influenced by its removal. Table 20 shows the results of the feature ablation analysis. According to the results in Table 20, the 'Distance to campus' feature does add some predictive value. However, this predictive value is very small.

		Accuracy
Validation set	**With 'Distance to campus'**	0.81
	Without 'Distance to campus'	0.80

Table 20: *Accuracies with and without 'Distance to campus' feature.*

In addition, the coefficient of the 'Distance to campus' feature was retrieved from the logistic regression model. It appeared that this coefficient is an extremely small positive number (0.000129). The small coefficient and the small added predictive value of the distance feature might indicate that the 'Distance to campus' feature does not play a crucial role in predicting student enrollment.

Moreover, the same analyzes have been done for the same distance subset but with only the potential students who live in the Netherlands. These analyzes also showed about the same results and that distance does not play a crucial role in predicting student enrollment.

6. Discussion

This chapter first provides a brief description of the research. Thereafter, the results of this research, as presented in chapter 5, are discussed and related to the existing literature. Moreover, the limitations of the current research are addressed and possible solutions are offered.

The goal of this study was to provide a weighing mechanism in which it can be defined which marketing efforts play a role for potential students in deciding to study at Tilburg University and what role the different marketing efforts play in that decision. To investigate the role of the different marketing events, logistic regression analyses were conducted based on attended marketing events. These logistic regression models were trained on 60% of the dataset and another 20% of the data was used to optimize the parameters of the models. Thereafter, the models with the tuned parameters were used on the other independent 20% of the data to test and evaluate the predictive power of the models. The coefficients from the logistic regression models were extracted to investigate the role of the different marketing efforts separately. In addition to the analyses on all education levels that are present at Tilburg University, separate analyses were conducted for potential bachelor and potential master students. This distinction was made to find out the differences in the decision-process between potential master and potential bachelor students.

According to the results of this research, a potential student that attends one of the marketing events of Tilburg University, can be predicted to be a prospective student or not with 83% certainty. When a distinction is made between potential bachelor and potential master students, the enrollment of potential master students (90%) can be predicted more accurate than potential bachelor students (77%). Although the accuracies of the majority classifier were already high, the accuracies of the logistic regression models are even a little higher.

By extracting the coefficients from the logistic regression models, it appears that the various marketing events can have positive or negative coefficient. To dive deeper into the results of the feature coefficients, the different marketing events will be discussed separately in the upcoming paragraphs following the offline-online distinction previously made in chapter 2. For each marketing event, it is described whether a student who attends that marketing event is more or less likely to enroll in a degree at Tilburg University, in which place it is in the

ranking (ranking can be found in Table 10) and what the predictive value of that marketing event is.

In this research, three different online marketing efforts were investigated. The results of this study show that potential students who download online brochures and/or watch online lectures and videos about the university are less likely to enroll in a degree at Tilburg University. This result might be explained by the fact that these two marketing events require only a limited amount of effort of the potential student. The small effort that the student wants to take to gather information can reflect the student's low interest. In addition, it could be the case that the online brochure and online videos do have a negative impact on the enrollment of students. According to Briggs (2006) and Moogan and Baron (2003) the possible negative impact of those two online marketing events might be due to incomplete, not detailed enough information and difficult language which leads to less attention from the reader. Looking at all education levels, the brochure is ranked number 7 and the online videos are ranked number 5 out of 11. These two online marketing events did only have small predictive value for the enrollment of potential master students and no predictive value for potential bachelor students. This might be explained by the fact that there are more potential master students who request a brochure and watch videos online than potential bachelor students.

The third online marketing event is providing the possibility to ask questions online. Students who attend this marketing event are, in contrast to the online brochure and online videos, more likely to enroll in a degree at Tilburg University than students who did not attend this marketing event. To ask online questions, more effort is needed from the potential students which might explain the more positive relation. In addition, it might be the case that the possibility to ask questions online has a positive influence on the enrollment of students. This possible positive impact could then be explained by the fact that the potential students like to be treated uniquely and receive almost instantaneous answers to their questions (Omboi & Mutali, 2011). However, this marketing event is ranked number 10 out of 11 in the ranking of importance. In addition, providing students the possibility to ask online questions has some predictive value for the enrollment of potential bachelor and potential master students. This result is in line with Johnston (2010) and Moogan et al. (1999) who state that being able to ask questions to a university representative is quite important in estimating whether a student enrolls at the university or not.

In addition to the three online marketing events, eight offline marketing efforts have been investigated. Two of these offline marketing events are the international education fair and the national education fair. The results of this research show that students who tend to visit the national education fair and/or the international education fair are less likely to enroll in a degree at Tilburg University. Except for potential master students who visit the national education fair, they are more likely to enroll in a degree at Tilburg University. The negative relation of the education fairs might be explained by the fact that multiple universities are present at an education fair. Potential students might not attend an education fair to especially receive information about one university but to make comparisons between different universities and degrees. However, apparently the potential master students are better focused in their search and/or the master degrees are presented more attractively and appealing at the national education fair. According to the ranking created in this study, the international fair is ranked number 2 and the national fair has ranked number 8 out of 11. In addition, this research showed that the two education fairs only have predictive power for the enrollment of potential master students and not for students that are interested in other education levels. This might be explained by the fact that the two education fairs are generally more often visited by potential master students than by potential students of other education levels.

Another offline marketing event that was subject of investigation is the international visit. According to the results of this research, students who make an international visit are a little more likely to enroll in a degree at Tilburg University than students who did not make an international visit. This positive relation might be explained by the fact that when the effort is made to make an international visit, the potential student is likely to already be very interested in the university. This effort could reflect the interest of the potential student. It appeared from the ranking that the international visit has the weakest relation with the enrollment of potential students, it is ranked number 11 out of 11. In addition, the international visit has no predictive value for any of the three created logistic regression models. The low ranking and lack of predictive value might be explained by the limited number of appearances in the dataset.

The fourth offline marketing event that was investigated in this research is the open lecture for potential bachelor students. According to the results of this research, students who

visit an open lecture are less likely to enroll in a bachelor degree than potential students who did not visit an open lecture. This negative relation can be explained in several ways. One possible explanation might be that when visiting an open lecture, the potential student attends only one lecture of one subject, which might not be representative for a whole study. Another explanation might be that at an open lecture the attractive parts of studying like the campus, the city and the student life do not get much attention. When looking at the ranking as created in this study, the open lecture has ranking 6 out of 11 and is has a relatively moderate relation with the enrollment of potential students. The open lecture for potential bachelor students does not have any predictive value which might be due to the low attendance rates.

Another two of the eight offline marketing events that were investigated in this research were the bachelor taster day and the master taster day. The results of this study show that potential students who tend to visit a bachelor taster day or a master taster day are more likely to enroll in a degree at Tilburg University. The positive relation of the taster days is in line with research of Mooney and Robben (1997) who state that a direct experience with the product leads to a more favorable evaluation of the product. In addition, in contrast to the open lecture, the taster days do provide a more representative view of studying at Tilburg University by not only visiting a lecture. For all education levels, the bachelor taster day has ranking 9 and the master taster day has ranking 1. It appears from this research that the master taster day also is of significant importance for potential bachelor students. This might be explained by the fact that potential bachelor students want to know what they can do after their bachelor degree or by the fact that they missed the bachelor taster day and still wanted to participate at a taster day at the university. In addition, the taster days do have some small predictive value in predicting whether a potential student becomes a confirmed student or not.

The last two offline marketing events that were addressed in this research are the bachelor and master open day. The results of this research show that potential students who visit the bachelor open day or the master open day are more likely to enroll in a degree at Tilburg University than potential students who did not attend these marketing events. For all education levels together, the bachelor open day gets ranking 4 and the master open day gets ranking 3. However, what is remarkable about the results is that the bachelor open day gets ranking 7 for potential bachelor students and ranking 3 for potential master students.

Moreover, the master open day gets ranking 2 for potential bachelor students and ranking 4 for potential master students.

Although the results concerning the open days seem odd, this are the results that were retrieved from the CRM dataset. The dataset does contain a reasonable number of cases in which a potential student attends a master open day and subsequently enrolls for a bachelor degree at the university. Likewise, the dataset does contain a number of cases in which a potential student attends only a bachelor open day and subsequently enrolls for a master degree at Tilburg University. The fact that some potential students that are interested in a bachelor degree visit a master open day might be due to the interest in what they can do after a particular bachelor or the impossibility to come to the bachelor open day. The fact that some potential master students visited the bachelor open day might be because they came for a bachelor several years ago but decided to study for their bachelor degree elsewhere and now come back to Tilburg for their master degree. Another explanation might be that the student achieved their bachelor degree at Tilburg University and decided, without visiting a marketing event, to also enroll for their master degree in Tilburg. Since this outcome differs from logical reasoning, it might be a reason for the marketing department of Tilburg University to consider their data saving and gathering and see if there is an error in the system.

Overall, it can be concluded from the results of this research that potential students who attend the marketing events that require effort from the potential students are more likely to enroll in a degree at Tilburg University than students who did not put much effort in their information gathering process. This finding is in line with research of Beatty and Smith (1987) and Liang (2012) who state that the amount of effort a potential student takes to gather information reflects the amount of engagement to the product. It can also be concluded that the marketing events where a student physically visits the campus, have more positive relations with student enrollment than marketing events where this is not the case. This result is in line with research of Pampaloni (2010) and Moogan, Baron and Harris (1999) who also state that physically visiting a school weighs heavily in the decision-making process to enroll in that university. In addition, this research shows that face-to-face contact generally appears to be of more positive influence on the enrollment of potential students than online marketing efforts. According to Herr, Kardes and Kim (1991) this might be since information retrieved in a face-to-face conversation is more accessible and has a profound influence on judgements.

At last, it can be concluded that university-specific marketing events do have a more positive relation with student enrollment than marketing events where more universities are present.

In addition to the eleven investigated marketing events, an extra feature was investigated: distance from place of residence to the Tilburg University campus. The results of this research show that distance to campus does have a very small positive coefficient that is almost equal to zero. In addition, the distance to campus did not have much predictive value. This result is not in line with research of Briggs (2006) and Kim and Gasman (2011) who state that the distance from residence to campus has an impact on the decision to enroll at that particular university. However, these analyses were performed on only a small part of the data because this information was not available for all potential students

The coefficients retrieved in this research proved to be very stable in the performed bootstrap analyses. In contrast with the ranking studies of de Jager and du Plooy (2010), Simões and Soares (2010), Pampaloni (2010) and Johnston (2010) the results of this research are more trustworthy and stable because there is no influence of the human memory and the data used in this research contained more and not only confirmed students. In addition, this research provides, in contrast to those four studies, insights in the direction of the relation of the different investigated marketing events and student enrollment. Therefore, this study is an enrichment to the existing literature. In addition, this research might be of practical value for, not only the department of Tilburg University, but also for other higher education institutes. Although the results of the research might be only generalizable to a certain extent because of the limited amount of investigated marketing events, the method used in this research can also be used in studies at other universities.

It should be noted that the accuracies of the tuned logistic regression models of this research were not much higher than the accuracies retrieved with the majority classifier. In order to make the models perform even better, more information about the potential students should be collected and analyzed. Examples of features that could add information and therefore make the models perform better are: age, gender, faculty of interest and number of family members who studied at Tilburg University. In addition, it might be that if more data would be available and the balance between the international and national students is known and representative for the whole university, the result of the distance to campus features might be different. This might be interesting for further research. At last, it is important to note that the results of this research only partly answer the research questions.

More investigation is needed to be able to fully describe the influence and impact of the different marketing events on student enrollment. However, this research is a step in the right direction and a good starting point for further research.

7. Conclusion

The goal of this research was to create a weighing mechanism in which it can be defined which marketing efforts play a role in deciding to study at Tilburg University and what role the different marketing efforts play in that decision. This weighing mechanism was created by means of logistic regression models from which the feature coefficients could be extracted. In the next paragraphs of this chapter the corresponding research questions, as formulated in chapter 1, are answered to some extend and suggestions for further research and recommendations for the marketing department of Tilburg University are given.

7.1. Answers to research questions

Q1: *How can student enrollment be predicted based on attended marketing events**?***

This first research question is a general question and represents the starting point of this research. According to this research, student enrollment can be predicted based on attended marketing events using a logistic regression model. For this research, the logistic regression model was chosen to be able to answer the upcoming three research questions but it might be possible that other algorithms are suitable to predict enrollment as well.

In this research, the logistic regression models were trained on 60% of the data and tuned on 20% of the data. Thereafter, the performance of the optimal logistic regression models were determined on the test set. In this research, it was possible to predict whether a lead in the dataset enrolls for a degree at Tilburg University or not with 83% accuracy.

Q2: *How are the different marketing efforts affecting the decision of potential students to enroll in a degree at Tilburg University?*

The results of this research show that the different marketing efforts can either have a positive or negative weight. To partly answer this research question, the coefficients of the different marketing efforts extracted from the model, based on all the available data, are discussed. These results are presented in order of a ranking of importance. The marketing effort which is ranked number 1 has the most important relation with the enrollment of potential students and the marketing effort which is ranked number 11 the least. This ranking can be seen in Table 10.

The marketing event which is ranked number 1, thus the most important marketing event, is the master taster day. A potential student who visits the master taster day is more

likely to enroll in a degree at Tilburg University than a potential student who did not visit this marketing event. The marketing event that is ranked number 2 out of 11 is the international education fair. Potential students who tend to visit the international education fair are less likely to attend Tilburg University than potential students who did not attend the international fair. The master open day is ranked number 3 out of 11. Potential students who tend to visit the master open day are more likely to attend Tilburg University than potential students who did not attend the master open day. Number 4 on the ranking is the bachelor open, which also has a positive relation with the enrollment of potential students. The online videos and lectures are ranked number 5 out of 11. Potential students who tend to watch online videos and lectures are less likely to attend Tilburg University than potential students who not tend to watch online videos. Then, at number 6, the open lecture for potential bachelors is ranked. This open lecture for potential bachelor has a negative relation with the enrollment of potential students. Thereafter, the online brochure is ranked on number 7 out of 11. Potential students who tend to request an online brochure are less likely to enroll in a degree at Tilburg University than potential students who did not request an online brochure. Number 8 in the ranking is the national education fair, which has a small negative relation with the decision to enroll in a degree at Tilburg University. The bachelor taster day is ranked number 9 out of 11. Potential students who tend to visit bachelor taster days are more likely to attend Tilburg University than potential students who did not visit this marketing event. Number 10 in the ranking of influence is the possibility to ask online questions. Potential students who tend to ask questions online are a little more likely to attend Tilburg University than students who did not ask questions online. The least important marketing event for the enrollment of students is the international visit. Potential students who did make an international visit are a little more likely to attend Tilburg University than students who did not made an international visit.

Overall, the results of this research show that potential students who attend the marketing events that require effort from the potential students are more likely to enroll in a degree at Tilburg University than students who only put a small amount of effort in their information gathering process. In addition, the results of this research show that events where a student physically visits the campus and has face-to-face contact, are more positively related to student enrollment than marketing events where this is not the case. At last, it can be

concluded that university-specific marketing events do have a more positive relation with student enrollment than marketing events where more universities are present.

Q3: *Which marketing efforts are most important and which marketing efforts the least in predicting student enrollment?*

To investigate the predictive value of the different marketing events for predicting student enrollment, an ablation analysis was performed over the model containing potential students of all education levels. The results of this research indicate that the different marketing efforts separately do only have a little predictive power for student enrollment. According to the results, the master open day has the most predictive power followed by the online brochure, national education fair, bachelor open day and the taster days for potential bachelor and master students. The possibility to ask online questions, the international education fair, the international visit, the open lecture for potential bachelor students and the online videos do, according to this research, not have any predictive power for the prediction task.

Q4: *What are the differences in weights of the marketing efforts in the decision to enroll between potential bachelor and potential master students?*

According to this research, there are some differences in weights for potential bachelor and potential master students. In other words, there are differences in the importance and direction of marketing efforts between potential bachelor and potential master students.

According to this research, potential students who tend to visit the national education fair are less likely to become a bachelor student of Tilburg University but more likely to become a master student of Tilburg University than potential students who did not attend the national education fair. In addition, potential students who tend to visit the open lecture are only less likely to become a bachelor student of Tilburg University than potential students who did not attend the national education fair. The last difference between the weights of the potential bachelor and potential master students is the bachelor taster day. Although the bachelor taster day has a positive relation with the enrollment of potential bachelor students, it has a negative relation with the enrollment of potential master students.

Moreover, there are also differences in the importance of the different marketing events between potential bachelor and potential master students. This study shows that online videos and the bachelor open day are of less importance for potential bachelor

students than for potential master students. In addition, research shows that both the master open day and the master taster day are of more importance for the enrollment of potential bachelor students than for the enrollment of potential master students.

7.2. Recommendations

7.2.1. For further research

Because there is currently no consensus in the literature about the importance of the different weights, future research could conduct a likewise study on a different university and compare the results with this study. In that way, it could be investigated if the relation of the different marketing efforts with student enrollment is similar over the different universities. In addition, if those results appear to be not comparable, one can look where these differences come from. To be able to investigate the real reasons for the different coefficients, a complementary qualitative study could be conducted. In addition, future research could focus on conducting experiments in which the real influence of the different marketing events on attitude and enrollment is investigated. These experiments could for instance contain a pretest and posttest.

Moreover, this research did not look at the impact on the weights of the different marketing events when a potential student attends a marketing event more than once. It might for instance be the case, that visiting a marketing event multiple times results in a different weight for that marketing event. This might also be interesting for future research. In addition, this research only focused on the controllable marketing events. Future research could also try to involve more characteristic features or non-controllable and partly controllable marketing efforts in the analyses. Adding more of these features might even result in a higher improvement of the model performance. At last, the distance feature could be reinvestigated because the results of this research and literature do not match.

7.2.2. For the marketing department of Tilburg University

The weighing mechanism created in this study might be useful for the marketing department of Tilburg University. When looking at the weights, it is advised to give more attention to online videos, brochures, open lectures for potential bachelor students, national education

fairs and international education fairs. These five marketing events appear to have a possible negative impact on the decision to start studying at Tilburg University.

In order to understand why these marketing efforts might have a negative influence, the marketing department should reflect on the events. For example, through online surveys in which the opinion and experience of potential students is asked after they visited one of the marketing events. By getting complementary information from the participants, the marketing department can gain insights into the reasoning behind the coefficients. In addition, by gaining these insights, the marketing department of Tilburg University is better able to change the content of the marketing events to the needs of the potential students.

In addition, to make further research at the marketing department at Tilburg University more efficient, the process of data collection should be modified. The data of the potential and confirmed students should either be stored in one dataset or in two separates but with a matching identifier between the two sets. Moreover, the creation date of the lead should be part of the dataset to be able to extract age from the birthdate or the potential students should be asked for age instead of their date of birth. By including the creation date of the lead, possible errors in the data could be prevented. For instance, all count features could be reset to zero when someone does not enroll within a certain time frame. This resetting would prevent creating odd leads in the dataset. When for instance a potential student visits a bachelor open day in 2012, a master open day in 2015 and enrolls for a master in 2015, you would only want to retain the master open day to prevent from odd results. In addition, more structure in the data could be added by using predetermined labels and less manually entered text.

References

Abrahamson, E. (2000). Change without pain. *Harvard business review*, *78*(4), 75-81.

Al-Fattal, A. (2010). *Understanding student choice of university and marketing strategies in Syrian private higher education*. University of Leeds.

Beatty, S. E., & Smith, S. M. (1987). External search effort: An investigation across several product categories. *Journal of consumer research*, *14*(1), 83-95.

Blackwell, R. D., Miniard, P. W., & Engel, J. F. (2001). Consumer behavior 9th. *South-Western Thomas Learning. Mason, OH*.

Blackwell, R., Miniard, P., & Engel, J. (2006). Customer behavior. *Mason: Thompson*.

Bloch, P. H., Sherrell, D. L., & Ridgway, N. M. (1986). Consumer search: An extended framework. *Journal of consumer research*, *13*(1), 119-126.

Bock, D. E., Poole, S. M., & Joseph, M. (2014). Does branding impact student recruitment: a critical evaluation. *Journal of Marketing for Higher Education*, *24*(1), 11-21.

Briggs, S. (2006). An exploratory study of the factors influencing undergraduate student choice: the case of higher education in Scotland. *Studies in Higher Education*, *31*(6), 705-722.

Briggs, S., & Wilson, A. (2007). Which university? A study of the influence of cost and information factors on Scottish undergraduate choice. *Journal of Higher Education Policy and Management*, *29*(1), 57-72.

Brown, C., Varley, P., and Pal, J. (2009), University course selection and services marketing. *Marketing Intelligence & Planning*, *27*(3), 310-325.

Cabrera, A. F., & La Nasa, S. M. (2000). Understanding the college-choice process. *New directions for institutional research*, *2000*(107), 5-22.

Coccari, R. L., & Javalgi, R. G. (1995). Analysis of students' needs in selecting a college or university in a changing environment. *Journal of Marketing for Higher Education*, *6*(2), 27-40.

Connor, H., Pearson, R., Pollard, E., Tyers, C., & Willison, R. (2001). The right choice? A follow up to making the right choice. London: Institute for Employment Studies/Committee of Vice-Chancellors and Principals.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5), 352-359.

Drummond, G. (2004). Consumer confusion: Reduction strategies in higher education. *International Journal of Educational Management*, *18*(5), 317-323.

Fischbach, R. (2006). Assessing the Impact of University Open House Activities. *College Student Journal*, *40*(1).

Gu, B., Park, J., & Konana, P. (2012). Research note- the impact of external word-of-mouth sources on retailer sales of high-involvement products. *Information Systems Research*, *23*(1), 182-196.

Herr, P. M., Kardes, F. R., & Kim, J. (1991). Effects of word-of-mouth and product-attribute information on persuasion: An accessibility-diagnosticity perspective. *Journal of consumer research*, *17*(4), 454-462.

Hemsley-Brown, J., & Oplatka, I. (2006). Universities in a competitive global marketplace: A systematic review of the literature on higher education marketing. *International Journal of Public Sector Management*, *19*(4), 316-338.

de Jager, J., & du Plooy, T. (2010). Information sources used to select a higher education institution: Evidence from south African students. *Business Education & Accreditation*, *2*(1), 61-75.

Johnston, T. C. (2010). Who and what influences choice of university? Student and university perceptions. *American Journal of Business Education*, *3*(10), 15.

Laurent, G., & Kapferer, J. N. (1985). Measuring consumer involvement profiles. *Journal of marketing research*, 41-53.

Liang, Y. P. (2012). The relationship between consumer product involvement, product knowledge and impulsive buying behavior. *Procedia-Social and Behavioral Sciences*, *57*, 325-330.

Kim, J. K., & Gasman, M. (2011). In Search of a “Good” College: First and Second Generation Asian American Students Describe Their College Choice Process. *Journal of College Student Development*, *52*(6), 706.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

Maringe, F. (2006). University and course choice: Implications for positioning, recruitment and marketing. *International Journal of Educational Management*, *20*(6), 466-479.

Maringe, F., & Mourad, M. (2012). Marketing for Higher Education in Developing Countries: emphases and omissions. *Journal of Marketing for Higher Education*, *22*(1), 1-9.

Meza, C. C. (2012). *Feiten en cijfers: De Nederlandse Universiteiten.* Den Haag: Rathenau Instituut.

Moogan, Y. J., & Baron, S. (2003). An analysis of student characteristics within the student decision making process. *Journal of Further and Higher Education*, *27*(3), 271-287.

Moogan, Y., Baron, S., and Harris, K. (1999), Decision-making behaviour of potential higher education students. *Higher Education Quarterly*, *53*(3), 211-28.

Moony, S., & Robben, H. (1997). Managing consumers' new product evaluation through direct product experience. In *the 26th EMAC Conference: Marketing: Progress, Prospects, & Perspectives*.

Obermeit, K. (2012). Students' choice of universities in Germany: structure, factors and information sources used. *Journal of Marketing for Higher Education*, *22*(2), 206-230.

Omboi, B. M., & Mutali, N. I. J. (2011). Effect of selected marketing communication tools on student enrolment in private universities in Kenya.

Pampaloni, A. M. (2010). The influence of organizational image on college selection: what students seek in institutions of higher education. *Journal of Marketing for Higher Education*, *20*(1), 19-48.

Provost, F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, 1-3.

Rowan-Kenyon, H. T., Bell, A. D., & Perna, L. W. (2008). Contextual influences on parental involvement in college going: Variations by socioeconomic class. *The Journal of Higher Education*, *79*(5), 564-586.

Schmidt, J. B., & Spreng, R. A. (1996). A proposed model of external consumer information search. *Journal of the academy of Marketing Science*, *24*(3), 246-256.

Schwartz, B. (2000). Self-determination: The tyranny of freedom. *American psychologist*, *55*(1), 79.

Simões, C., & Soares, A. M. (2010). Applying to higher education: information sources and choice factors. *Studies in Higher Education*, *35*(4), 371-389.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. *AI 2006: Advances in Artificial Intelligence*, 1015-1021.

Veloutsou, C., Lewis, J. W., & Paton, R. A. (2004). University selection: information requirements and importance. *International Journal of Educational Management*, *18*(3), 160-171.

Veloutsou, C., Paton, R. A., & Lewis, J. (2005). Consultation and reliability of information sources pertaining to university selection: some questions answered?. *International Journal of Educational Management*, *19*(4), 279-291.

Whitehead, J. M., Raffan, J., & Deaney, R. (2006). University choice: What influences the decisions of academically successful post-16 students?. *Higher Education Quarterly*, *60*(1), 4-26.

Appendix A: description of features

Feature name	Description	Values
Level	Education level of interest	1 = Master 2 = Bachelor 3 = Research Master 4 = PhD 5 = Pre-Bachelor 6 = Pre-Master
Converted	Registered in StudieLink	0 = No 1 = Yes
Brochure.Count	Count how often a brochure has been requested	Numerical value
International.Fair.Count	Count how often an international fair has been visited	Numerical value
Open.Lecture.Ba.Count	Count how often a Bachelor open lecture has been visited	Numerical value
Taster.Day.Ba.Count	Count how often a Bachelor taster day has been visited	Numerical value
Taster.Day.Ma.Count	Count how often a Master taster day has been visited	Numerical value
Education.Fair.Count	Count how often an education fair has been visited	Numerical value
Open.Day.Ba.Count	Count how often a Bachelor open day has been visited	Numerical value
Open.Day.Ma.Count	Count how often a Master open day has been visited	Numerical value
Webinar.Count	Count how often an online lecture or informative video has been viewed	Numerical value
International.Visit.Count	Count how often an international visit has been made	Numerical value
Online.Question.Count	Count how often a question has been asked online by the student	Numerical value