An Analysis of Non-Contractual Churn in the B2B Hotel Industry

De Wit, L.S.

ANR: 837191
U-Number: 1255441

Tilburg University

Master thesis: Data Science: Business and Governance 2017

First Reader: dr.ing. S.C.J. Bakkes
Second Reader: Assistant Professor A.T. Hendrickson

Date: Friday, July 28, 2017

Preface

This thesis was made possible by several people and organizations. I would like to make use of the possibility to thank these people and organizations:

I would, first of all, like to thank my internal supervisor Sander Bakkes for reviewing my work and his support throughout the process of writing the thesis. I would also like to thank the second reader, Drew Hendrickson, for evaluating my thesis. Next, I want to thank Tilburg University (and the Data Science: Business and Governance teaching staff in particular) for teaching me how to individually investigate scientific topics related to the analysis of 'Big Data'.

Second, I want to thank my external supervisor Tekla de Veer, as well as Albert Blom and Bas Karsemeijer for helping me with the technical parts of my thesis and pointing me in the right direction. I would also like to thank Yourzine B.V. as a whole for granting me the opportunity to write my thesis on their behalf.

Third, I would like to thank Wouter van de Loerakker and his colleagues for granting me access to the Bilderberg database and answering my context specific questions.

Laurens de Wit
Tilburg, July 2017

Abstract

The current study examines the potential of analysing churn in a context in which it has not been studied before: the hotel industry. To narrow down the investigation, the context is specified to being B2B and non-contractual. Also, two research questions are constructed that ask (1) whether a standard model or a context-specific expansion thereof is needed to analyse churn in this context, and (2) what predictors are of the highest importance in this context. Data are obtained from a hotel CRM database. By utilizing three predictive methods (logistic regression, SVM, and AdaBoost), and evaluating these methods with ROC AUC scores, it can be established that a standard churn prediction model (in combination with a logistic regression) is optimal. Furthermore, it appears that several predictors are of importance with 'recency' being the most prominent. Next, it is shown that the ROC AUC score of the optimal model/algorithm combination indicates that the model performs better than the established baseline. Also, by constructing a cumulative gain chart on the basis of the optimal model/algorithm combination, practical relevance can be established (potentially reaching up to 30% more churners when half of the database is targeted in a retention campaign). Therefore, the conclusion is drawn that churn analysis is applicable to the non-contractual B2B hotel industry.

Keywords: Churn, B2B, Non-contractual, Hotel Industry, Data Mining

Contents

Preface……………………………………………………………………………………………………1

Abstract…………………………………………………………………………………………………..2

Contents…………………………………………………………………………………………………. 3

1. Introduction…………………………………………………………...…………………….…………5

1.1. Hotel Industry CRM Data Mining: Churn……………...………………………………………5

1.2. Mining Real Hotel Data: Bilderberg……………………………………………………...7

1.3. Problem Statement………………………………………………...…………………...8

1.4. Outline………………………………………………………………………………………9

2. Literature Overview……………………………………………………………………………….…10

2.1. Developments in Churn Analysis…………………………………………………………….…10

2.1.1. Inter-Industrial Churn Analysis…………………………………………………….…..11

2.1.2. Specifying Churning Context…………………………………………………….…..12

2.2. Time Window……………………………………………………………………………….…..13

2.3. Methods to Analyse Churn…………………………………………………………………….…..14

2.3.1. Attributes to Analyse Churn……………………………………………………......14

2.3.2. Methods to Predict Churn (Predictive Analyses)………………………………………..15

2.3.3. Methods to Describe Churn (Descriptive Analyses)……………………………….17

2.4. Contributions of the Current Study………………………………………………………...17

3. Method……………………………………………………………………………………………19

3.1. Description of Dataset………………………………………………………………………..19

3.1.1. Attributes…………………………………………………………………………..20

3.1.2. Time Window and Churn Operationalisation……………………………………...20

3.1.3. Creating Unique Identifier Variable per Company………………………………...21

3.2. Feature Creation and Selection……………………………………………………………….22

3.2.1. Selecting and Creating Features for the Standard Model…………………………..22

3.2.2. Selecting and Creating Features for the Bilderberg Model………………………...25

3.3. Preparing the Models for Analysis…………………………………………………………30

3.3.1. Missing Values………………………………………………………………….30

3.3.2. Biased Data Entries……………………………………………………………..30

3.3.3. Filters……………………………………………………………………………30

3.3.4. Standardization………………………………………………………………….31

3.4. Splitting Data into Training and Test Sets……………………………………………………31

3.5. Predictive Analysis………………………………………………………………………...32

3.5.1. Logistic Regression……………………………………………………………..33
3.5.2. Support Vector Machine…………………………………………………………...33
3.5.3. Boosted Decision Tree……………………………………………………………..34
3.6. Evaluation Techniques…………………………………………………………………….35
3.6.1. ROC AUC……………………………………………………………………….35
3.6.2. Cumulative Gain Chart…………………………………………………………….36
3.7. Descriptive Analysis……………………………………………………………………….37
3.8. Answering Research Questions and Addressing Problem Statement…………………………38
4. Results……………………………………………………………………………………………..39
4.1. Comparing Models (RQ1)………………………………………………………………....39
4.1.1. Results of Parameter Optimization: Logistic regression…………………………...40
4.1.2. Results of Parameter Optimization: Support Vector Machine……………………..40
4.1.3. Results of Parameter Optimization: Boosted Decision Tree………………………40
4.1.4. Summary of Experimental Findings: RQ 1………………………………………...41
4.2. Feature Analysis (RQ2)…………………………………………………………………... 42
4.2.1. Descriptive Results: Logistic Regression…………………………………………42
4.2.2. Descriptive Results: Support Vector Machine……………………………………..44
4.2.3. Descriptive Results: Boosted Decision Tree………………………………………45
4.2.4. Summary of Experimental Findings: RQ 2………………………………………...46
4.3. Applicability of Model………………………………………………………………………47
5. Discussion……………..……………………………………………………………………….. 48
5.1. Answering Research Questions………………………………………………………………48
5.1.1. Answering RQ 1…………………………………………………………………...49
5.1.2. Answering RQ 2…………………………………………………………………...50
5.2. Addressing the Problem Statement……………………..…………………………………51
5.3. Contributions of Study…………………………………...……………………………...52
5.4. Direction for Future Research…..………………...……………………………………..53
References…………………………………………………………………………………………55
Appendices………………………………………………………………………………………...59
Appendix 1: Action Orders Matrix…………………………………………………………..59
Appendix 2: Company Specific Matrix……………………………………………………...61
Appendix 3: Text Manipulations……………………………………………………………62
Appendix 4: Features Used in General Model………………………………………………63
Appendix 5: Features Used in Bilderberg Model……………………………………………64

1. Introduction

1.1. Hotel Industry CRM Data Mining: Churn

The hotel industry, which is part of the hospitality industry, is an industry with a significant amount of direct contact with customers. As such, the application of customer relationship management (CRM) is seen as playing a crucial role in the hotel business (Wu & Lu, 2012). However, as Mohammed and Rashid (2012) note, there still is a marked lack of CRM research in the hotel industry, even though CRM research in said industry can be utilized to improve customer value, boost customer satisfaction, and obtain larger amounts of revenue (Wu & Lu, 2012). Especially in combination with data mining techniques, CRM related analyses can tremendously influence factors such as the ones mentioned above (Ngai, Xiu, & Chau, 2009).

Since the early 2000s, researchers have been investigating the hotel industry with data mining techniques such as predictive modelling and clustering (Weatherford & Kimes, 2003). The usage of these techniques is made possible by centralized reservation systems, which allow for a large accumulation of customer data, organized in databases (Magnini, Honeycutt, & Hodge, 2003). Properly utilizing such databases can help to formulate marketing strategies and maximize profits (Magnini et al., 2003). A data mining technique in the context of the current study is defined as training a learning algorithm on historical data and subsequently predicting output for unseen data in the future. The actual learning is achieved by statistically mapping input features of the (historical) training data, after which certain combinations of features in unseen (future) data can be translated into output (Van Leemput, 2016).

The focus of CRM data mining in the hospitality industry thus far has been on topics such as customer profile development through clustering (Min, Min, & Eman, 2002), forecasting of guest arrivals as well as occupying rates of rooms through predictive modelling (Weatherford & Kimes, 2003), and service quality by the use of sentiment analysis (Duan, Cao, Yu, & Levy, 2013). However, as stated above, CRM research within the hospitality industry is still far from complete. In the domain of data mining specifically, there appears to be a lack of research in which predictive models are utilized that deal with the propensity of customers to attrite, or churn. Churn, in this context, means that a customer made purchases at some point in the past, and made no subsequent purchases during an established period in the future (see definition 2). The current study is designed to fill this gap in the literature, by investigating the possibilities of predicting churn in the hotel industry. In order to further stress the importance of churn modelling in the hotel industry, factors such as customer loyalty and customer retention, as well as their potential positive effects on profits (Shoemaker & Lewis, 1999) will be discussed in section 2 (literature overview).

In classic churn research, the customer is assumed to have a routine relationship with the company, such as the payment of monthly telephone bills (Mozer, Wolniewicz, Grimes, Johnson, & Kaushansky, 2000; Verbeke Dejaeger Martens, Hur, & Baesens, 2012; Van Leemput, 2016). It is speculated that churn prediction research is lacking within the hotel industry due to the non-routine nature of the relationship between customer and hotel. One reason for the non-routine nature of this relationship, is the fact that, although certain hotels have incorporated customer loyalty programs such as Frequent Guest Programs (FGP's), the relationship between customer and hotel is non-contractual in most cases (Tepeci, 1999). However, Hoppman and Thede (2005) have demonstrated that successfully predicting churn is also possible in a non-contractual setting. In order to create a more routine nature of the non-contractual relationship between customer and hotel, the current study is designed to investigate business-to-business (hereafter B2B) churn rather than business-to-consumer (hereafter B2C) churn. Naturally, a B2C relationship refers to when a company does business with the end customer, and a B2B relationship refers to when a company does business with other companies (Murphy, 2007). This focus should make churn prediction more viable, seeing as, generally speaking, a B2B relationship is of a more routine nature than a B2C relationship (Claycomb, Iyer, & Germain, 2005). The difference between the two churning contexts and the influence of either a contractual- or non-contractual relationship has been studied thoroughly and will be addressed in section 2. Only selecting customers with a non-contractual B2B relationship leads to a restriction of the input that will be used, which can be translated into the following definition of a customer:

Definition 1: "A **customer** included in the current study is only considered to be as such, if a business context is apparent and if there is no formal contract that binds the relationship between customer and hotel."

To define churn, the definition of churn posed by Jahromi, Stakhovych, and Ewing (2014) will be used. These authors stress the difficulty of defining churn in a non-contractual setting, for one needs to take into account very specifically what is meant by "churn" in this context. Jahromi et al. (2014) further elaborate that non-contractual churn should not be seen as having a permanent nature, which mostly is the case for contractual churn. Rather, churn should be seen as the inactivity of a customer in a certain time period when the customer was observed as active in an earlier time period. Therefore, the definition of a churner is formulated as follows:

Definition 2: “A customer is considered to be a **churner** when proximate earlier activity is not followed by subsequent activity.”

The selection of a time period, and other criteria such as the frequency of activity within a time period, will be further elaborated on and revisited in section 3 (methodological overview).

1.2. Mining Real Hotel Data: Bilderberg

To explore B2B churn in the hotel industry, the current study will investigate data accumulated by Bilderberg Hotel Group between 2013 and 2016 by the means of a case study. Bilderberg is one of the largest hotel companies active in The Netherlands, with 17 offices distributed throughout the country and a revenue of 70 million euros in 2016 (Reijden, 2016). Bilderberg has several tools to engage the B2B market such as an extensive amount of conference halls and rooms. However, competition in the Dutch hotel industry is significant, and being able to properly identify churn will likely improve Bilderberg’s market position. B2B Churn statistics are high when compared to frequently investigated industries such as telecommunications (Van Leemput, 2016), but considered to be representative of the B2B hotel industry. Possible explanations for these differences are offered in section 2.

The dataset supplied by Bilderberg will be mined by using data mining techniques as these were described in subsection 1.1. In this sense, churn can be identified before it has arisen and important predictors of churn can be established. As will be explained in section 2 and 3, relevant data mining techniques (or algorithms) for the current study are a logistic regression, a support vector machine, and a boosted decision tree. Results of these algorithms will be obtained by cross-validating their area under the receiver operating characteristic to optimize parameters on a training set, after which another area under the receiver operating characteristic can be obtained by running the optimized algorithms on a test set. Finally, a cumulative gain chart will be constructed on the basis of the best performing algorithm to establish practical relevance (the reader is, again, referred to sections 2 and 3 for further elaboration on these terms).

As stated above, constructing a churn model for Bilderberg, as well as determining which factors are of importance in the process of constructing this model, is expected to be beneficial for the hotel concern. In related work, Jahromi et al. (2014) investigated an online retailer with a B2B context, and predicted profit would significantly increase if their model would actively be incorporated by the retailer through a customer retention campaign, that was designed to specifically target churners. It is expected that if similar results are obtained with the current study, Bilderberg will be given an extra opportunity to solidify their market position and increase profits. Other practical examples of successful churn prediction and further elaboration of the need to timely identify churn are offered in section 2. Furthermore, if Bilderberg is given insight into how they can influence the churn rate through which predictors are considered statistically relevant, it is speculated that Bilderberg will be able to target certain 'churn enhancing' factors specifically. In this sense, a profile of a potential churner can be constructed. Ultimately, this knowledge is expected to further decrease churn rates and therefore increase profits.

1.3. Problem Statement

As was explained in section 1.1, there seems to be a need for more CRM research in the hotel industry, and more specifically, CRM research that utilizes data mining techniques. To contribute to closing this gap in the literature, the current study is designed to investigate the possibility of constructing a non-contractual B2B churn prediction model that is applicable to the hotel industry. The problem statement of this thesis is formulated in a way that formally summarizes the abovementioned exploration.

Problem Statement: *To what extent is the construction of a churn prediction model viable in the non-contractual B2B hotel industry?*

The problem statement will be addressed by studying the case outlined in section 1.2 through the use of the data mining techniques mentioned in the previous subsection (and further elaborated on in sections 2 and 3), and by exploring earlier research that focussed on analysing churn in different settings (section 2). To investigate what this type of churn prediction model should look like, as well as which features are considered most important for Bilderberg to take into consideration, the following two research questions are formulated:

Research question 1: *Is the extension of a standard churn prediction model by incorporating context-specific features useful when analysing churn in the non-contractual B2B hotel industry?*

The first research question will be answered by constructing a generally applicable churn prediction model based on predictor variables[1] with a high (previously proven, see Jahromi et al., 2014) predictive power (hereafter 'standard' model), and comparing this model with an extended version that also contains

[1] Note that the terms predictor variables, features, and variables will be used interchangeably throughout the thesis.

more Bilderberg database specific predictor variables (hereafter ‘Bilderberg’ model). As stated above, further elaboration with regard to the differences between these models is given in sections 2 and 3.

Research question 2: *Which features are considered most important in order to predict churn in the non-contractual B2B hotel industry?*

In order to answer the second research question several descriptive techniques will be used. These techniques are linked to the data mining techniques mentioned earlier, and will be discussed in-depth in sections 2 and 3. Eventually, a hierarchical order of feature importance will be constructed, in combination with an overview of the direction of the effects.

1.4. Outline

The current study is structured as follows. In section 2, a literature overview of what has been accomplished in the field of churn prediction thus far will be offered. Section 3 offers an in-depth description of the methods that will be used to address the problem statement and answer the research questions. In this section, pre-processing of the raw data as these were supplied by Bilderberg will be elaborated on, after which the construction of the predictive- and descriptive analyses used will be explained. In section 4, the best performing model is chosen based on several evaluation techniques (RQ 1), and a hierarchical order of feature relevance will be presented (RQ 2). Section 4 will be wrapped up by addressing the applicability of the optimal model/algorithm combination. In section 5 the problem statement and research questions will be discussed further. Section 5 will be concluded by the offering of recommendations for subsequent research.

2. Literature Overview

In this section, relevant churn literature will be discussed. By a review of what churn analysis has focused on in the past, lessons can be learned from other industries. In this way, the research domain can be further specified and narrowed down. Moreover, by identifying similar studies, a couple of predictor variables can be identified in advance, and context specific predictive- and descriptive churn analysis techniques can be chosen. To start, general developments in churn modelling will be explored in subsection 2.1, and arguments that demonstrate the importance of these models will be given. The churn modelling developments will be tracked through an inter-industrial view in subsection 2.1.1, and further narrowed down to the non-contractual B2B domain of churn research in subsection 2.1.2. Next, in subsection 2.2, the importance of the selection of a proper time window will be emphasized, after which the state of the art of churn analysis methods will be elaborated on in subsection 2.3. The latter is further divided by an overview of relevant churn predictor variables in subsection 2.3.1, relevant predictive methods in subsection 2.3.2, and relevant descriptive methods in subsection 2.3.3. The literature overview section will be concluded in subsection 2.4, by offering a short explanation of how the current study will add new findings to the existing field of churn analysis.

2.1. Developments in Churn Analysis

Markets have become increasingly saturated in several industries, causing businesses to shift their focus from customer acquisition to customer retention. This choice is supported by findings that acquiring and securing new customers is becoming increasingly expensive and difficult (Hadden, Tiwari, Roy, & Ruta, 2007). Churn prediction is seen as one of the most effective tools of this new CRM-based approach (Hadden et al., 2007). As stated in the introduction, churn becomes apparent when a customer that has been doing business with a company since a certain point in the past, ceases to do further business with the company in the future. If a company is affected by churn, a company can be blemished through declining levels of profit, the loss of a substantial amount of price premium, and a reduction in inter-customer referrals (Reichfeld & Sasser, 1990). An example of what can be achieved by accurately anticipating churn is given by Tamaddoni Jahromi, Sepehri, Teimourpour, and Choobdar (2010). These authors summarize literature in which they found that an increase of 5% in customer retention rate can increase the average net present value of customers by 35% for software companies and 95% for advertising agencies.

Tammaddoni Jahromi et al. (2010) further explain that two basic approaches exist to retain customers. The first approach is known as a ‘blanket-campaign’ and basically targets the entire customer base through low cost techniques such as emailing. The second approach is more tailored to individual customers, and aims to provide those customers who are likely to churn with direct incentives not to do so. An example of such an incentive for the hotel industry would be to contact high probability churners with discounts on bookings in the future. Naturally, the more individually tailored approach is possible only through churn analysis and is generally more effective than the blanket campaign. Tammaddoni Jahromi et al. (2010) split the more targeted approach further by establishing two subcategories. The first subcategory entails a reactive approach that deals with customer churn at the moment it happens. In the current context, the reactive approach is not viable as there is no possibility for Bilderberg to know exactly when churning happens (there is no contract to terminate). The second subcategory entails a proactive approach which attempts to identify specific churn cases before they even happen, and proactively offers retention incentives to these specific customers. The current study is designed to potentially realize retention campaigns that fall under the tailored proactive approach of customer retention management.

2.1.1. Inter-industrial Churn Analysis

In relatively recent research, churn analysis has proven its worth in several industries such as banking (Popović & Bašić, 2009), telecommunications (Verbeke et al., 2012), grocery retail (Miguéis, Van den Poel, Camanho, & Cunha, 2012), and E-Commerce (Jahromi et al., 2014). As stated in the introduction, churn analysis seems to be lacking in the hotel industry. To understand churn in this specific industry, one will have to compare it with churn in other industries. From a database perspective, industry comparison is possible because, in theory, any industry that keeps digital records of its customers should have access to basic transactional data (Buckinx, Verstraeten, & Van den Poel, 2007). As will be explained in section 2.2, using basic transactional indicators such as recency, frequency and monetary value has been proven to lead to the development of well-functioning churn prediction models (Jahromi et al., 2014), and therefore, these types of indicators will be used to construct the standard model (RQ 1). Furthermore, both the grocery industry, as well the hotel industry, employ customer loyalty programs (Bolton, Kannan, & Bramlett, 2000; Meyer-Waarden & Benavent, 2009). These programs entail the potential for customers to become loyal. Subsequently, when a customer is considered to be a loyal customer, churn analysis becomes more useful. Therefore, the existence of customer loyalty programs is one of several factors that validates churn research in these industries.

However, it should be noted that each industry is unique, and that industries can only be compared to a certain extent. For example, the telecommunications industry is based on a contract between customer and company (Verbeke et al., 2012), whereas in retail the relationship between customer and company is mostly non-contractual (Jahromi et al., 2014). Likewise, reasons to churn across industries tend to differ significantly. Within the telecommunications industry churn mostly arises because customers find better prices at a competitor, though within the banking industry churn is observed most frequently when service is perceived as being better at a competitor (Popović & Bašić, 2009; Verbeke et al., 2012). Possible explanations for churn in the hotel industry will be further elaborated on in the next subsection. Another inter-industrial difference is that churn rates throughout, and even within industries, tend to differ. For example, in the telecommunications study done by Verbeke et al. (2012), the churn rate of the different telecommunication case studies varied between 1% and 10%. This churn rate is relatively low when compared to the churn rate in the grocery retail study done by Miguéis et al. (2012) which was 44%. A relatively high churn rate is also expected for Bilderberg, and would seem logical, as consumption in the hotel industry is generally of a less routine nature than for example consumption in the telecommunications industry. However, as noted in the introduction, the current study is designed to incorporate routine bookings as much as possible by only investigating B2B churn. This choice was made under the assumption that B2B bookings in the hotel industry are of a more routine nature than B2C bookings in the hotel industry (Claycomb et al., 2005).

2.1.2. Specifying Churning Context

Churn research in the B2B domain is relatively new and tends to differ from churn research in the B2C domain (Jahromi et al., 2014). Because the current research will focus on the B2B domain, a disquisition on the important differences between B2B and B2C churn analysis will be given. Jahromi et al. (2014) explain that the availability of B2B 'Big Data' is more limited when compared to the abundance of data in the B2C domain. Furthermore, practices to exploit data and transform it into information are still underdeveloped in B2B companies. Also, predictor variables that are available and useful for churn analysis tend to differ between the domains. For the hotel industry this difference is apparent in the availability of sociodemographic indicators such as age. Anil Kumar and Ravi (2008) have found that age can be a useful predictor with respect to predicting churn in a B2C setting. It is speculated that, for Bilderberg specifically, age could also be interesting in a B2C analysis, because older customers are generally more loyal than younger ones in the service industry (Caruana, 2002). However, as B2B bookings are generally not made by the guest himself, the guest lacks the influence on the booking that is responsible for the importance of sociodemographic features such as age. Nonetheless, as explained above, choosing to exclusively investigate the B2B customer base of Bilderberg is expected to increase the validity of the current study due to the relatively high loyalty levels of B2B customers (compared to B2C customers) in the hotel industry.

Another specification in the current research design that needs to be addressed is the focus on non-contractual churn. Traditionally, to analyse churn, research has focused on a relationship between business and customer that is bound through a contract (i.e. Verbeke et al., 2012). This kind of churn, generally, is of a more permanent nature. As was discussed in the introduction, non-contractual churn cannot be defined as being permanent (Jahromi et al., 2014), seeing as even when customers appear not to be active during a certain period, chances are that they will be during subsequent periods. Because customers with a non-contractual relationship do not actively end the relationship, as is common in contractual settings, churn in non-contractual settings is often an 'invisible phenomenon' (Tammadoni Jahromi et al., 2010). Due to this invisibility, Bilderberg has experienced trouble identifying loyal customers that cannot expressively be labelled as being loyal through, for example, participating in loyalty programs. Through a non-contractual churn analysis such customers may reasonably be expected to be easier to identify. The operationalisation of non-contractual churn in the current context is further elaborated on in the next paragraph.

2.2. Time Window

Because defining non-contractual churn as being a permanent phenomenon is incorrect (see subsection 1.1, Jahromi et al., 2014), the current study sees churn as having a temporary nature. Churn will be investigated in a time window that is predetermined. A predetermined time window in this sense means that customer data will be utilized to determine the probability of activity of customers in a period in the future, based on a period in the past. In order to do so, a time window is established by splitting the data in two periods. These periods are known as the calibration period and the prediction period[2] respectively (Buckinx & Van den Poel, 2005; Tamaddoni, Stakhovych, & Ewing, 2016). In this way, a customer is considered to be a churner when he has made purchases in the calibration period and no subsequent purchase(s) in the prediction period.

Developing a calibration and prediction set to analyse churn is not trivial and subject to obstacles. One of these obstacles is the duration of the prediction period under investigation. In their comparative study on customer churn, Tamaddoni et al. (2016) explain that the duration of the prediction period "should be set in a way that it (1) captures activity/inactivity of customers with a fairly long interpurchase time and (2) captures defection of those with a short average interpurchase time as soon as possible." These authors further elaborate that, in order to achieve these notions, they set the duration of the prediction period in their study according to two steps: "(1) First, we sort customers in ascending order based on their average

[2] The reader should be aware that splitting the data into a calibration and prediction period differs from splitting the data into a training and test set. An elaboration on the differences between these two splits is given in subsection 3.4.

interpurchase times and (2) the prediction period is set to be approximately equal to the average interpurchase time of the last customer in 99% mass of the sorted customer base." In order to reduce the risk of using prediction periods that are too long (and not being able to identify churn in time), as well as the risk of using prediction periods that are too short (and incorrectly labelling customers with comparatively large interpurchase times), the time window in the current study is designed to mirror the time window as described in the study by Tamaddoni et al. (2016).

2.3. Methods to Analyse Churn

Research focusing on churn analysis can be split into descriptive and predictive analyses. The former is designed to find and elaborate on factors that explain churn, whereas the latter is designed to identify potential churners before actual churn becomes apparent (Tammadoni Jahromi et al., 2010). Descriptive analyses are popular throughout marketing literature, whereas predictive analyses are popular throughout data mining literature (Ahn, Han, & Lee, 2006; Fathian, Hoseinpoor, & Minaei-Bidgoli, 2016). Research questions 1 and 2, as they were introduced in the opening section, can be investigated through the use of predictive and descriptive analyses respectively. The first research question is related to whether a standard model (based on only a couple predictors with previously, in other contexts, found high predictive power) is sufficient to accurately predict churn in the current context, or whether a so called Bilderberg model (one that expands the standard model with several context-specific predictor variables) is needed. The second research question explores which predictor variables, of either the standard model or the Bilderberg model (depending on which model is considered best), are most valuable when it comes to making this prediction. This subsection will provide the reader with a theoretical backing for the predictive- and the descriptive analysis respectively. However, as the predictors used in these analyses are identical, a short theoretical framework on the choice of these predictor variables will follow first.

2.3.1. Attributes to Analyse Churn

As briefly mentioned in the opening section, the current study makes use of basic transactional predictors such as recency, frequency and monetary value, in order to construct a standard churn prediction model. Relevant theory on this model will be elaborated on next. Theoretical backing for the Bilderberg model (the extension of the standard model with more industry-specific predictor variables) will not be supplied here because, as explained in the introduction, churn analysis in the hotel industry has not been investigated before. This means that there are no hotel-industry-specific variables available in churn literature. To argue why such a model was used anyway, promising findings from an exploratory data analysis on the predictor variables that were used to construct the Bilderberg model are presented in section 3. The reader will now be supplied with an overview of what predictor variables might be relevant to include in the standard model. This information is based on churn research that has focussed on identifying predictor variables that easily generalize to other contexts (or industries).

Jahromi et al. (2014) explain in their study on non-contracutal B2B churn analysis (E-Commerce) that a churn prediction model can be kept simple and applicable to most settings, by employing only a limited set of predictors with maximum predictive power. In order to do so, these researchers picked recency, frequency and monetary value as predictor variables. As the construction of the standard model in the current study has similar interests, comparable predictor variables to the ones used by Jahromi et al. (2014) will be incorporated in this model. In this sense, the standard model can be seen as a baseline model, proven to be successful in the past in other industries. By also constructing the Bilderberg model (which is an expansion of the standard model), an assessment can be done as to whether the standard model is adequate in the current context as well. The directions of the effects of the predictor variables mentioned above are described next.

A recent study by Coussement and De Bock (2013) on churn analysis in the online gambling industry, describes the important role of recency and frequency variables in predicting customer churn. These researchers have found that when a customer's purchase is more recent, the probability of activity over the next period increases. Furthermore, Coussement and De Bock (2013) found that the higher a customer's purchasing frequency, the lower the probability for the customer to churn becomes. Monetary value has not been investigated as recently as the other two variables. However, it has been demonstrated in the past that a relationship between churn and monetary variables exists (Schmittlein & Peterson, 1994).

2.3.2. Methods to Predict Churn (Predictive Analyses)

In a recent study that compares churn prediction from a methodological point of view, Tamaddoni et al. (2016) distinguish between probability models and data mining models as the two overarching categories in academia. Tamaddoni et al. (2016) explain that probability models simply make use of probability distributions to model the behaviour of customers, whereas data mining models analyse large amounts of data to find patterns. In the managerial domain, and thus from a more practical point of view, Tamaddoni et al. (2016) stress the importance of the RFM model[3]. The RFM model is frequently used as a segmentation technique and consists of three measures (recency, frequency and monetary), which are

[3] Note that the RFM model has served as the inspiration of including predictor variables such as recency, frequency, and monetary value in churn prediction modelling with data mining techniques, but that employing an RFM model is not the same as employing a data mining model.

combined into a three-digit RFM cell code, covering five equal quintile (Wei, Lin, & Wu, 2010). Tamaddoni et al. (2016), through data simulation, found that the models under investigation (probability models, data mining models, and RFM) performed differently as they varied the churn ratio. According to these researchers, data mining techniques are preferred especially when the churn ratio increases. As the case under investigation is subject to a high churn ratio, the employment of data mining techniques is preferable.

A priori choosing the best data mining techniques is difficult, because it is expected that techniques tested and proven to be best in different studies might be context dependent. Therefore, the selection of data mining techniques will be based on adequate performance of these techniques across several different industries. Lacking in the current literature is a comprehensive overview that compares churn from an inter-industrial point of view. Therefore, the current study will base the choosing of appropriate data mining techniques on the results of three methodologically comparative studies conducted across several industries. To relate to the current research design as much as possible, the studies considered for this purpose were limited to research conducted in the non-contractual B2B domain. Moreover, studies of a more recent date were considered to be of higher value than older studies due to the fast pace of innovation in data mining and corresponding techniques (Larose, 2014). Therefore, studies were only taken into consideration when they were conducted in 2014 or later. A final note with regard to study selection is related to the variables under investigation. In all three of the chosen studies recency, frequency and monetary value play an important role in explaining churn, indicating another link to the current study (remember that the standard model will be constructed with these types of predictor variables).

The first study reviewed is the churn prediction research conducted by Tamaddoni et al. (2014), which has already been thoroughly discussed above due to its relevance to the current study in content and design. As stated before, these researchers investigated churn prediction in a non-contractual B2B setting with E-Commerce as context. Predictive techniques used in this study were two types of decision trees (simple and cost-sensitive), logistic regression, and a boosting model (improving decision trees with an ensemble learner method). Evaluation was based on the area under the receiver operating characteristic and cumulative lift measures, and indicated that boosting significantly outperformed the decision trees, but only marginally outperformed logistic regression. Further work in non-contractual B2B E-commerce was done by Gordini & Veglio (2016) who compared a support vector machine based on the area under the receiver operating characteristic parameter-selection technique (SVMauc) with a neural network, a classic support vector machine and logistic regression. Based on accuracy, the area under the receiver operating characteristic, and top-decile lift these authors found that the SVMauc outperformed the other models. The final study considered here focused on non-contractual B2B churn in the logistics industry (Keramati, Ghaneei, & Mirmohammadi, 2016). These researchers predicted churn by comparing the performance of a decision tree classifier, a multilayer perceptron, a support vector machine with a poly kernel as well as an RBF kernel, and a simple logistic regression. Based on accuracy, precision, recall, and the F1 measure, Keramati et al. (2016) found that the decision tree significantly outperformed the other classifiers.

In all these studies, logistic regression was picked as a benchmark technique due to its popularity in marketing literature. Tamaddoni et al. (2014) also found that their best model only marginally outperformed logistic regression. Therefore, logistic regression will be incorporated in the current study. As Keramati et al. (2016) found satisfying results for their decision tree classifier and Tamaddoni et al. (2014) found that boosting a decision tree beat their simple decision tree, a boosted decision tree classifier will also be incorporated in the current study. Finally, due to its successful performance in the study conducted by Gordini and Veglio (2016), a support vector machine will be included in the current study. More information on these three classifiers, parameter selection techniques, and evaluation metrics will be offered in section 3.

2.3.3. Methods to Describe Churn (Descriptive Analyses)

To describe churn, or in other words, to find the most important predictor variables and the direction of their effects, several techniques are described throughout the relevant literature. It was decided to utilize three different types of descriptive techniques, one for each predictive algorithm. For the logistic regression and the boosted decision tree methods, descriptive analyses that are specific for these techniques will be used (Pedregosa et al., 2011, see section 3 and 4). For the support vector machine no such method exists (for an explanation why see section 3). Therefore, ablation will be used to assess feature importance of the support vector machine model. Ablation entails running a predictive algorithm on the dataset with all the predictors to establish an error rate. Next, the same analysis is ran, but with one (category of a) predictor left out. This process is repeated for each (category of a) predictor. In this way, predictor importance can be ranked by descending error enlargements (for an example from churn literature see Pudipeddi, Akoglu, & Tong, 2014).

2.4. Contributions of the Current Study

From a practical point of view, churn analysis is one of the most powerful tools of data driven CRM. A properly constructed churn model can significantly increase profits of the company under investigation

(Tammaddoni, 2014). Therefore, churn prediction and description can be a valuable asset to Bilderberg specifically, and the hotel industry in general, if proven that churn analysis is applicable to this context.

From an academic point of view, no studies exist in churn literature that have been conducted in this unique setting yet. The current study is designed to fill this gap in the literature, as well as reduce the general lack of data-driven CRM research in the hotel industry. Relevant literature as described in this section will be used as a guideline for operationalizing churn in the hotel industry, constructing valuable predictor variables, and building a predictive- as well as a descriptive analysis. Eventually, an indication can be given of to what extent the construction of a churn prediction model is viable in the non-contractual B2B hotel industry.

3. Method

In this section, a methodological overview of the study will be given. Again, this section will commence with an overview of each subsection[4]. In subsection 3.1, a technical description of the dataset will be offered that covers the different attributes that are part of it. Furthermore, to operationalize the outcome variable (churn), the construction of the time window that will be used to split the data into a calibration- and prediction period is explained more practically. The first subsection will conclude by explaining how the identifier variable, which was needed to label each unique company, was created by reducing several data points (bookings) to one unique vector per company. In order to answer the first research question (which asked whether a standard model needed to be extended to predict churn in the current context), the construction of the standard- and Bilderberg models are discussed in subsection 3.2. This will be achieved by explaining how the predictor variables belonging to each model were operationalized. In subsection 3.3, several pre-processing steps taken to prepare the dataset for analysis are elaborated on. These pre-processing steps are taken to reduce bias and to incorporate context information. Pre-processing includes the discussion of missing values, biased data entries, filters used, and standardization. In subsection 3.4, for evaluation purposes, choices related to splitting the data into a training and testing set will be clarified. In subsection 3.5, the predictive analyses chosen (logistic regression, support vector machine, boosted decision tree) will be discussed. These predictive analyses (in combination with the evaluation techniques discussed in subsection 3.6) are constructed to answer the first research question, because picking the best performing model will be based on the model that makes the best predictions. In subsection 3.6, evaluation techniques for the predictive analyses will be discussed, starting with the area under the receiver operating characteristic and ending with the cumulative gain chart. In subsection 3.7, in order to answer the second research question (that asks what features are of the highest importance when analysing churn in the non-contractual B2B hotel industry), the descriptive analyses chosen per predictive algorithm will be discussed. Finally, in subsection 3.8, a brief but comprehensive explanation will be offered on how exactly the research questions will be answered and the problem statement addressed in the next section (section 4, results), by using the methods as they were described in the current section.

3.1. Description of Dataset

In this subsection, a description of the dataset will be offered. The database of Bilderberg consists of matrices related to mail, website (click) and CRM data. For a churn analysis, the CRM part of the database is of primary interest. The CRM database consists of 937,580 data entries of which 401,073 have

[4] Note that this overview of the coming subsections, in order to inform the reader as adequately as possible, is more extensive than the overview supplied in the two previous sections. This decision was made to establish a red line to which the reader can refer back to while reading this section.

a B2B context. In subsection 3.1.1, an overview of attributes that exist in the database will be given. Subsection 3.1.2 offers a more technical description of the time window that was discussed in section 2.2, as well as the subsequent operationalisation of churn as the outcome variable. Finally, in subsection 3.1.3 an elaboration is offered on how each data entry was traced back to one unique booking, and how each unique booking was traced back to one unique company. All data manipulation and visualization described in this subsection and the next two subsections (3.2 and 3.3) was done with R by using the ggplot2, dplyr, lubridate, readr, and tidyr packages.

3.1.1. Attributes

The B2B CRM data of Bilderberg is distributed across two matrices, which contain all relevant attributes. These attributes will be manipulated to construct, in order of discussion, the outcome variable (churn), the identifier variable (the company), and the predictor variables (variables that contain information that can potentially point to whether a company is a churner or not). The first matrix contains all action orders and the second matrix is tailored to company specific data, based on the action orders in the first matrix. An overview of the attributes included in the first matrix is given in appendix 1, and an overview of the attributes included in the second matrix is given in appendix 2.

3.1.2. Time Window and Churn Operationalisation

As was elaborated on in section 2.2, the data will need to be split into a calibration and prediction period, because a distinction between a period of activity and a period of potential inactivity has to be made[5]. To split the data into these two periods, an appropriate point in time needs to be established that covers as much real churn as possible. As was explained in section 2.2, the creation of a time window is based on the methodological churn study done by Tamaddoni et al. (2016). These authors demonstrated that a time window can be established empirically. To achieve this, the average inter-purchase time per company is calculated for each company with a frequency of at least 2, ordered ascendingly, and then the split is set to be equal to the average inter-purchase time of the last customer in 99% mass of the sorted customer base. This process led to the finding of an ideal split of 861 days per period. However, as the data were registered between 12-03-2013 and 06-03-2017, the actual time period only covers 1455 days. This means that the maximum split can be set to approximately 727 days. Thus, it must be noted that not all non-churners could be identified fully correctly in the calibration set. However, as the difference between the ideal and actual split is relatively small, the actual split is considered to be adequate.

[5] Note that a figure is included in subsection 3.4 to inform the reader of the differences between the data splits of the calibration- and prediction period on the one hand and the training- and test set on the other hand.

With the chosen split between calibration and prediction periods, the target variable (churn) can be assigned to each unique company that has a transactional history in the calibration period. Hence a company is considered to be a churner (coded as 1) when the company has been active during the calibration period but not during the prediction period. A company is considered to be a non-churner (coded as 0) when the company has been active during the calibration period and has made subsequent purchases during the prediction period. Coding churn as 1 and non-churn as 0 was chosen intentionally to highlight the focus on churners (Tamaddoni, 2016). This operationalisation is in accordance with the definition of churn offered in section 1 (definition 2).

3.1.3. Creating Unique Identifier Variable per Company.

To properly study churn, each vector analysed must belong to one company. This is not straightforward, because the Bilderberg database consists of vectors that are unique for each booking and not for each company. As vectors are unique for each booking, data belonging to a single company is distributed across several different vectors. These vectors must be aggregated together for each unique company. In order to do so, a connection between the two matrices has to be made, several extra features have to be created (see subsection 3.2), and several text manipulations are necessary. This subsection will explain the steps that were taken to make sure that the information belonging to each unique company is summarized in each unique vector. For clarity’s sake, the reader is asked to keep in mind that the process described here merely concerns the creation of the identifier variable. The creation of the other features (which will serve as predictor variables in the analyses) will be discussed in the next subsection.

A link between the matrix containing all action orders and the matrix containing company specific data can be made through the variable ‘company_key’, which is an identifier variable present in both matrices (see appendix 1 and 2). This identifier represents all business related bookings by a company. It should, however, be noted that this identifier, due to poor administration, can still be different for the same company. When a company makes a booking, it can have multiple data entries under the same name if the company, for example, books on behalf of several employees. This means that one data entry could be part of several other entries that all belong to one company at the same point in time. These different entries of the same company on the same date had to be aggregated to properly establish certain predictor variables (see subsection 3.2). Furthermore, the variable ‘ses_session_type’ (containing the levels ‘conference’, ‘hotel’, and ‘dummy’) had to be used as a filter to remove dummy data entries. These data entries were ‘off-the-record’ purchases (for example when a guest had already checked out and ordered a drink at the bar), and contained bias in many cases.

Once data entries of the same ‘company_key’ on the same date were aggregated, the same was done for all unique ‘company_key’ entries. However, in order to accurately express differences between the calibration period and the prediction period, all data entries were split according to the time window that was established (see above). At this point, there were two different sets of data aggregated on the basis of the ‘company_key’ variable. Nevertheless, as there were still different data points of the same company in the same time period, aggregation was not complete yet. The next step was to aggregate the remaining data points on the basis of the ‘prd_full_name’ variable (the variable that contained the company names) for each time period separately. In order to do so, assumptions had to be made with regard to what companies belonged together. For example, a company could be represented as Rabobank B.V. at one time and Rabobank BV at another time, and therefore have different entries of the ‘company_key’ attribute. To deal with these deviations, several text manipulations such as excluding capital letters, spaces, and punctuation marks were performed. For an overview of these manipulations see appendix 3. Now that different data points of the same company shared the same name, these data points could be further aggregated leading to one unique data point, or vector, per unique company.

3.2. Feature Creation and Selection

Now that the identifier variable is established for each vector, the predictor variables can be created and selected. This concerns the variables that will be used to predict churn. Therefore, in this subsection, the creation and selection of features (predictor variables) that will be used in the analyses will be discussed. In order not to bias the calibration period (for example by taking the frequency of a company throughout the whole dataset), all feature creation will be performed twice, once for the calibration part of the data and once for the prediction part of the data (the two different sets that were created by splitting according to the time window). As the first research questions concerns the construction of two different models (standard and Bilderberg), this subsection is further divided in two subsections. First, features created and selected for the standard model will be discussed (subsection 3.2.1). Second, features created and selected for the Bilderberg model will be discussed (subsection 3.2.2).

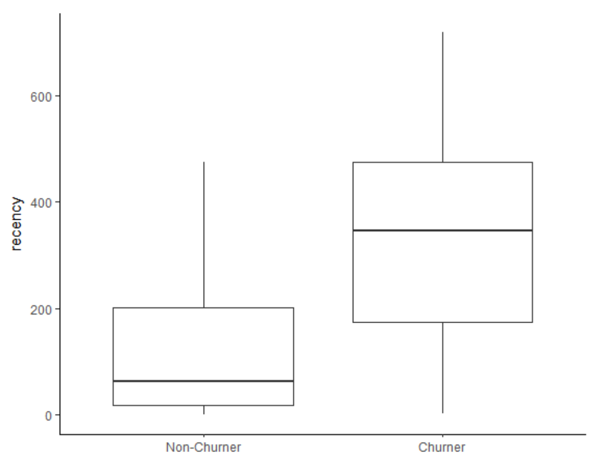
3.2.1. Selecting and Creating Features for the Standard Model

As stated in section 2, the standard model is built purely with features that can be found in any transactional database. These features are related to recency (and other simple variables associated with time), frequency, and monetary value, and are known to have a relatively high predictive power (see section 2).[6] Below, an overview of these features, and how they were created is offered. For a more concise overview of these features, the reader is referred to appendix 4.

It is expected that the more recent a company made purchases, the lower their propensity to churn becomes (Coussement & De Bock, 2013). This expectation can also be seen graphically in the Bilderberg data in figure 1. This figure shows that the biggest part of the churners, relative to the biggest part of the non-churners, is concentrated at higher values of 'recency'. This variable was created by simply subtracting the last date in the time period of interest (so, for example, the end of the calibration period) from the last date a company made a transaction for each unique company. In this sense, the higher the number of this feature was, the further away the company's most recent booking was from the end of the time period of interest.

Figure 1

The relation between churn and recency. Churners seem to be concentrated at higher values of recency, and non-churners seem to be concentrated at lower values of recency. Recency in days is depicted on the y-axis, and the relative spread of non-churners compared to churners is depicted on the x-axis.



Two other time related variables were created. The feature 'observation_period' was created to capture how long a company was active during a time period (so the period from the first activity until the end of the calibration period) and served both as a predictor of churn as well as a filtering method (see subsection 3.3). This feature was created in a similar way as 'recency'. Furthermore, the average interpurchase time per company, as was calculated in section 3.1.2, was also incorporated into the model (as 'interpurchase'), due to the favourable results of a study on churn done by Buckinx and Van den Poel

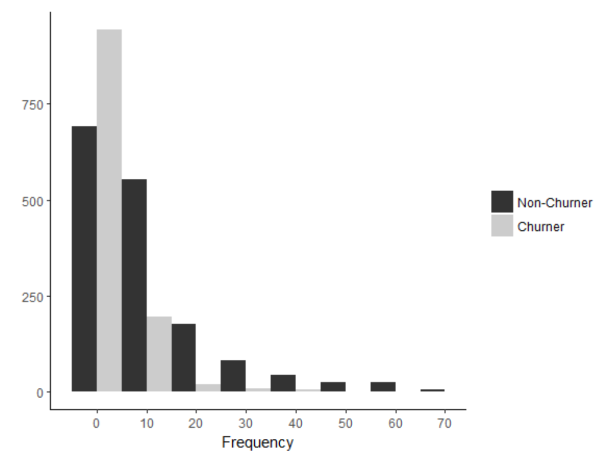
[6] The reader is reminded that the standard model, as it has been proven to be successful in other industries in the past, is seen as a baseline model.

(2005). These authors found that the lower the interpurchase time was per customer, the lower the probability of churning became.

Next, as it is expected that companies with more frequent purchases have a lower chance to churn (Coussement & De Bock, 2013), the ‘frequency’ variable was created. As can be seen in figure 2, the churn rate in the Bilderberg dataset indeed seems to decrease as values of frequency increase. These values were calculated by first aggregating all data that shared the same ‘company_key’ and ‘ses_datetime’ (booking date, see appendix 1 and 2). This was done because data entries with identical figures on these two attributes were the same company buying several products (i.e. rooms) during one session, which means that several data entries could be part of an activity that should possibly have a frequency of one. If these frequencies would have been counted separately, the feature would be biased. Next, the remaining data entries per company were counted by using the (cleaned version of the) ‘prd_full_name’ variable as identifier, which led to the creation of a column with these numbers.

Figure 2

The relation between churn and frequency of visits. As the numbers of visits per company increases, the chance of churning appears to decrease.

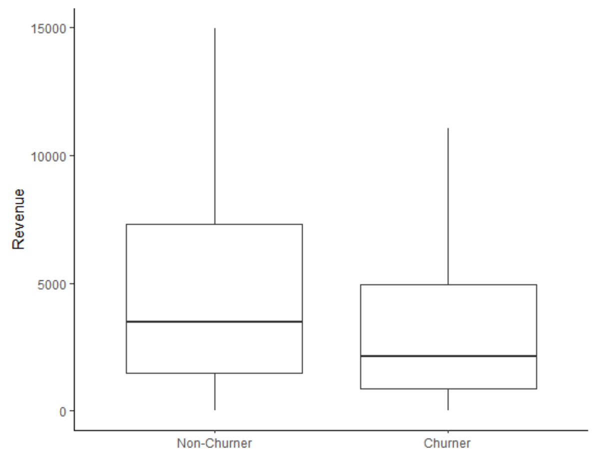


Note. Frequencies of 1 were filtered out of the data (see subsection 3.3.3), and frequencies above 70 (about 50 rows) were not includeded in this visualization for clarity’s sake.

To finalize the standard model, the monetary value feature was created. As can be seen in figure 3, it appears that companies with higher values of ‘ors_revenue’ are observed more frequently as being non-churners. With regard to monetary value, the ‘off-the-record’ dummy (one of the three ‘ses_session_type’ levels) bookings were important, as these included additional purchases of guests that were not registered as normal bookings. However, these bookings were only relevant for the food and beverage types of revenue (2, 3, and 4). Therefore, all dummy bookings and normal bookings that had a match in 'prd_full_name' were aggregated for 'ors_revenue_2', 'ors_revenue_3', and 'ors_revenue_4'. Then, by simply summing all nine types of revenue together for each data entry that was part of the same company (by using a combination of 'company_key' and 'prd_full_name' again), one feature entailing the monetary value per unique company could be created.

Figure 3

The relation between churn and total revenue. Churners seem to be concentrated at lower values of revenue, and non-churners seem to be concentrated at higher values of revenue. Total revenue per customer over the course of the calibration period in euros is depicted on the y-axis, and the relative spread of non-churners compared to churners is depicted on the x-axis.



3.2.2. Selecting and Creating Features for the Bilderberg Model

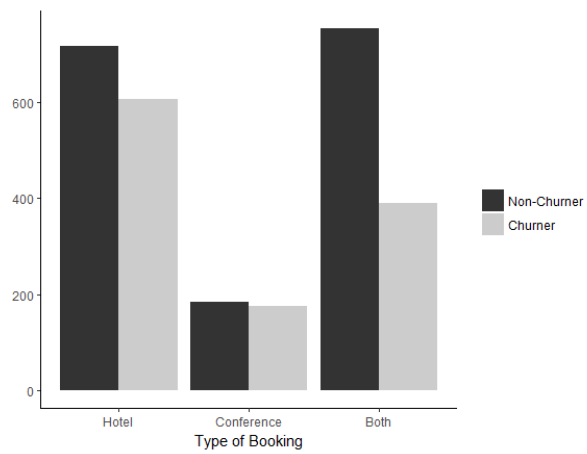
For the Bilderberg model, as it is an expansion of the standard model, the features discussed in the previous subsection will be incorporated into the model in combination with several context-specific features. Because the Bilderberg model is an expansion of the standard model, it is assumed that adding context-specific features should *significantly* increase performance in order to determine that the Bilderberg model is superior, and thus has performed better than the baseline imposed by the standard model. Otherwise, constructing a model that is more complex (more complex in the sense that is consists of more variables), would only unnecessarily complicate the model by, for example, a larger running time. Therefore, if the performance of the Bilderberg model appears to be similar or worse than the performance of the standard model, the standard model is considered to the best model. As was pointed out in section 2, to the best of our knowledge no hotel industry churn literature exists that can be used to motivate the choice of the extra features chosen for the Bilderberg model. Therefore, an overview of the features chosen will be given, supported by an exploratory data analysis. Where possible, theoretical backing from churn research in domains other than the hotel industry will also be offered. Again, for a more concise overview of these features, the reader is referred to appendix 5.

Feature(s) Related to Main Products

Firstly, the difference between the two main products Bilderberg offers was incorporated into the model by the creation of a feature based on the 'ses_session_type' attribute. This feature ('Type') indicated whether a company only had hotel bookings (coded as 0), conference bookings (coded as 1) or both (coded as 2). This distinction was made because it is expected that customers who purchase more than one product, generally have a reduced probability to churn (Buckinx & Van den Poel, 2005), which (as can be seen in figure 4) also appears to be the case for Bilderberg's customers. Furthermore, as can be seen in figure 4, churn rates in the Bilderberg dataset appear to be higher for companies that only purchase conference related bookings compared to companies that only purchase hotel related booking.

Figure 4

The relation between churn and type of booking. Churning appears to be much more prevalent among companies that only purchase one product (either a hotel booking or a conference booking) compared to companies that book both products.



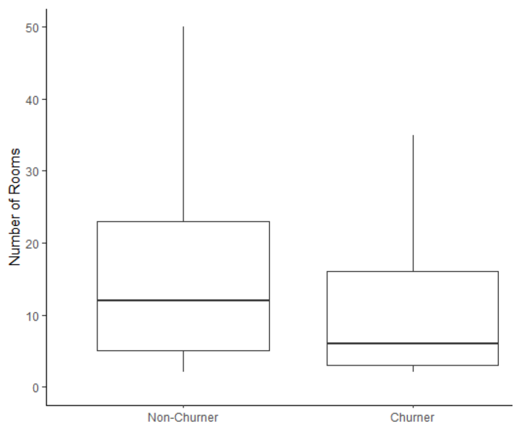
Feature(s) Related to Booking Specific Attributes

Secondly, the booking specific attributes 'ors_nr_of_rooms', 'ors_nr_of_adults', and 'ors_length_of_stay' (see appendix 1 and 2) were incorporated into the model by aggregating these numbers for each company. As can be seen in figure 5 A, B, and C the churn rate appears to decrease as the numbers of these attributes increase.

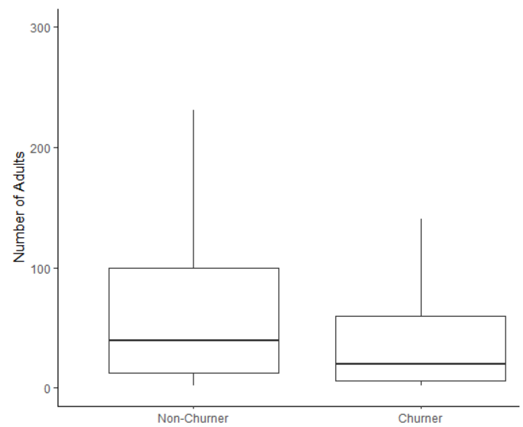
Figure 5

The relationship of churn with booking specific variables.

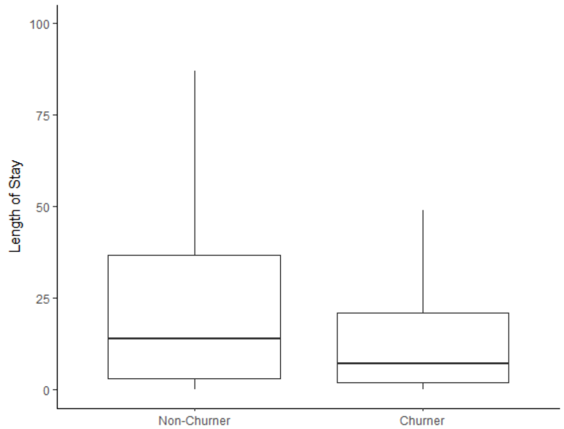
A: The relationship of churn with amount of rooms. Churning companies seem to book fewer rooms than non-churning companies. The total number of rooms booked during the calibration period in is depicted on the y-axis, and the relative spread of non-churners compared to churners is depicted on the x-axis.



B: The relationship of churn with amount of adults. Churning companies seem to book for fewer persons than non-churning companies. The total number of adults for whom rooms were booked during the calibration period is depicted on the y-axis, and the relative spread of non-churners compared to churners is depicted on the x-axis.



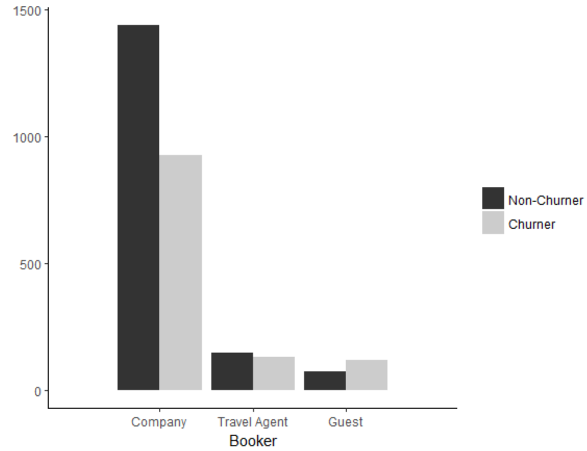
C: The relationship of churn with length of stay. Churning companies seem to stay shorter than non-churning companies. The total length of stay for each company during the calibration period is depicted on the y-axis, and the relative spread of non-churners compared to churners is depicted on the x-axis.



Also, as can be seen in figure 6 (on the next page), it appears that if the company makes the majority of the bookings itself, the chance to churn is lower than when a travel agent makes the majority of the bookings. Furthermore, when a guest makes the majority of the bookings himself (of course, still indicating that a business context is apparent) it appears that the chance to churn is higher than the chance not to churn. Therefore, the final booking specific attribute that was incorporated into the model was ‘booker’ (which indicated by what body most bookings were made). The creation of this variable was achieved by aggregating the amount of bookings done by the company itself, the guest himself, and travel agencies. Subsequently, the variable could be labelled as 0 if the company made the majority of the bookings itself, 1 if the travel agent made the majority of the bookings, and 2 if the guest made the majority of the bookings himself.

Figure 6

The relationship between churn and majority booker. If the company did most bookings, the chance to churn seems to be lower than the chance to churn if either the guest himself did most bookings or when the travel agency did most bookings.

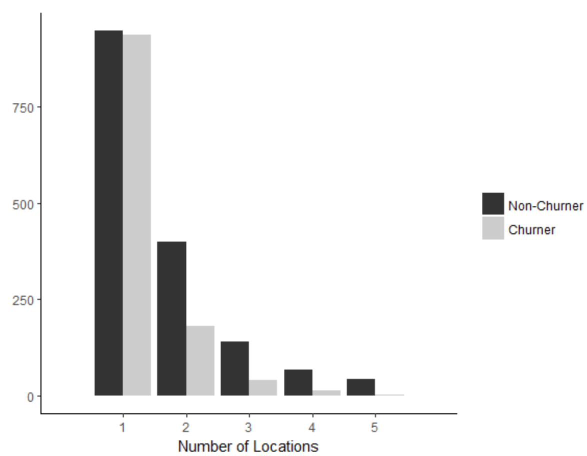


Feature(s) Related to Location

Thirdly and finally, based on the visualization in figure 7, the 17 different locations of Bilderberg were incorporated into the Bilderberg model. This was achieved by aggregating counts for the 'ses_property_key' variable for each unique booking per company. Then the amount of unique Bilderberg locations booked per company were counted and added as a feature ('nr_locations').

Figure 7

The relationship between churn and the amount of Bilderberg locations visited by the company. It appears that the more locations were visited by a company, the lower the chance of that company churning becomes.



3.3. Preparing the Models for Analysis

After identifying each unique company and assigning proper feature values to each company, several steps were taken to further prepare the data for analysis. As was stated in the opening paragraph of this section, these steps are necessary to deal with bias and incorporating context specific information.

3.3.1. Missing values

Missing values had to be removed from the dataset in order not to bias the findings. The only relevant missing values in the data were empty ‘prd_full_name’ entries. After aggregating all relevant attributes to the features mentioned above, this led to one data point with a missing identifier. As this aggregation was based on all ‘prd_full_name’ missing values, this row was expected to be biased (as it could have belonged to several different companies) and was deleted from the data. Furthermore, the column ‘ors_revenue_9’ was fully excluded from the Bilderberg model because all values in this column were equal to 0.

3.3.2. Biased Data Entries

Certain data entries of ‘prd_full_name’ appeared to be biased in the sense that they did not represent a single company. For example, all weddings that were celebrated at one of the Bilderberg hotels were included in the database as ‘WEDDING’. If these entries would have been aggregated together on the basis of the ‘prd_full_name’ variable, one company called ‘WEDDING’ would have been created. In order to avoid this bias, all data related to weddings, banqueting, world meetings, and Bilderberg itself were removed from the analysis. The same was done for bookings that were made under the name of a travel agency that was actually booking for a company.

3.3.3. Filters

As was mentioned in the opening paragraph of this section, filters are necessary to incorporate context information, and, again, deal with bias. In this case, context information means the information supplied by Bilderberg. Two filters, based on the features selected for analysis, were used (both for the standard as well as for the Bilderberg model). First, all companies that displayed no activity during the first half of the calibration/observation period were filtered out based on the ‘observation_period’ feature. This was done in order to avoid bias in frequency and recency figures (Tamaddoni & Stakhovych, 2016). Second, all companies that had a frequency of 1 were excluded from the analysis. This was done because the current study focuses on customers that have a long-term relationship with Bilderberg. Bilderberg does not consider B2B customers having a long-term relationship with the hotel when there has only been one visit during the calibration period. Because these filters exclude certain companies as customers, the definition of a customer (definition 1) as it was given in section 1 is refined:

Definition 1: "A **customer** included in the current study is only considered to be as such, if a business context is apparent, there is no formal contract that binds the relationship between customer and hotel, there were at least 2 visits during the calibration period, and the customer visited at least once during the first half of the calibration period."[7]

3.3.4. Standardization

As was elaborated on in section 2, the utilization of a support vector machine in order to investigate churn is highly supported throughout the relevant literature. Therefore, the current study will make use of this algorithm (see subsection 3.5.2). In order to use a support vector machine, data has to be standardized (Ben-Hur & Weston, 2010). Furthermore, standardizing data generally leads to more favourable results when it comes to predictive analyses (Vercellis, 2011). Standardization was done by setting the mean to 0 and the standard deviation to 1 (Z-scores).

After executing all the above-mentioned steps, the calibration set contained data on 2830 unique companies. The general model contained 6 dimensions and the Bilderberg model contained 12 dimensions. The final step that had to be taken (before analysing the data could commence) was to split the data into a training and test set.

3.4. Splitting Data into Training and Test Sets

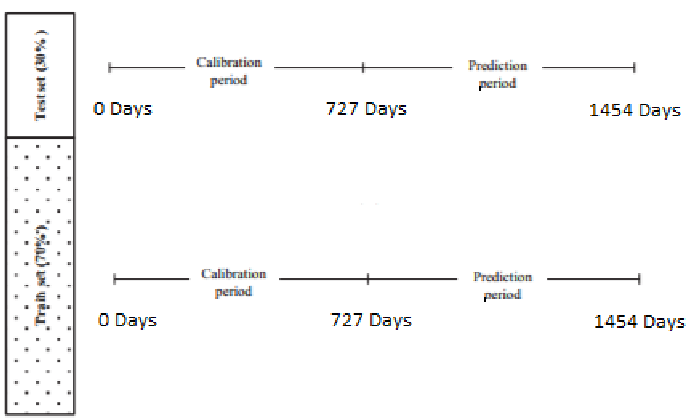
For evaluation purposes, the data was split into a training- and test set. The training set was created to train the models (standard and Bilderberg) on the data, and to tune the parameters of the algorithms used via cross-validation (see next subsection). Furthermore, by assessing which features were deemed most important during training, a hierarchical order of feature importance can be given (RQ 2, see subsection 3.7). By running the optimized versions of all model/algorithm combinations on the test set it can be decided which combination (and thus which model) predicts churn best (RQ 1).

[7] Note that refining this definition could only be done at this point, seeing as an in-depth explanation of the time window and filters were necessary. This choice was made in order not to confuse the reader in the opening section of thesis.

The training- and test sets were constructed in a balanced fashion, which means that the ratio churners/non-churners was kept equal (Van Leemput, 2016). The churning rate in both sets was approximately 41%. The training set consisted of 70% of the data and the test set covered the other 30% of the data. A 70/30 split was chosen as it is common in churn literature (for a recent example, see Keramati et al., 2016). After performing the split, 1981 data points remained to train the predictive algorithms. Although this amount of data points for training purposes might seem low, Tammadoni et al. (2016), in their study that concerned the investigation of churn from a methodological point of view, showed that churn prediction models are not sensitive to sample sizes as long as there are more than 1,000 observations. The reader is reminded of the earlier split of the data into a calibration- and prediction period according to the time window that was established in subsection 3.1.2. Figure 8 illustrates the difference between the calibration- and prediction period split on the one hand, and the training- and testing set split on the other.

Figure 8

This figure is meant to clarify the difference between two splits that were done. As can be seen, the calibration part of the data (with labels for churners and non-churners, as they were established by using the prediction period part of the data, included) takes up 70% of the training data and 30% of the test data. In this sense, models can be trained on the larger part of the (labelled) data, and their performance can be evaluated on a smaller (also labelled) part of the data.



3.5. Predictive Analysis

As was explained in subsection 3.1.2, the outcome variable (churn) in the current study is binary (0 = non-churner, 1 = churner), meaning that classification with probability assignment is in order. In this subsection the predictive analyses (or classifiers/algorithms) touched upon in subsection 2.3.2 will be discussed more thoroughly and technically[8]. The predictive analyses will be used to answer the first research question. After all, the model (standard or Bilderberg) that yields the best predictive results will be considered best. In the following sections, relevant parameters and settings are discussed for the algorithms that were chosen in subsection 2.3.2 (based on studying relevant churn literature). The chosen algorithms were: a logistic regression, a support vector machine, and a boosted decision tree. All analyses, both predictive and descriptive, will be done with Python by using SciKit Learn (hereafter sklearn) packages.

3.5.1. Logistic Regression

Logistic regression (hereafter LR) is a common classifying technique used in marketing research (Akinci, Kaynak, Atilgan, & Aksoy, 2007). It is popular because it is relatively easy to comprehend compared to other classifiers (Bucklin & Gupta, 1992), posterior probabilities are directly available, and its findings are generally efficient and robust (Neslin, Gupta, Kamakura, Lu, & Mason, 2006). Despite its name, LR is not a regression technique but a linear model for classification. Classification is performed by using a logistic function to obtain the probabilities of the outcome of a single vector. For the current research design, there is only one parameter of interest to tweak. This parameter is known as the C parameter and it specifies the strength of regularization, with smaller values indicating stronger regularization. For more technical information on how this classifier works and how it is implemented in an applied setting, the reader is referred to the comprehensive work of Hosmer Jr., Lemeshow, and Sturdivant (2013).

3.5.2. Support Vector Machine

Although less common than LR, a support vector machine (hereafter SVM) is a popular predictive technique to use for prediction problems with a churning context (for a recent example see Gordini & Veglio, 2016). SVMs create what is called a hyperplane in a high-dimensional space to classify entities either as class A or class B (or churners and non-churners). Gordini and Veglio (2016) explain that this is achieved by maximizing the margin between the two hyperplanes separating the two classes. These authors, by analysing in depth work done by Kim and Sohn (2010) and Kim, Shin, and Park (2005), further explain that SVMs have the following advantages: “(1) there are only two free parameters to be chosen, namely, the upper bound and the kernel parameter; (2) the solution of SVMs is unique, optimal, and global since the training of an SVM is done by solving a linearly constrained quadratic problem; (3) SVMs are based on the structural risk minimization principle, which means that this type of classifier

[8] The reader is reminded that a theoretical backing as to why the predictive algorithms, discussed in subsections 3.5.1 up to and including 3.5.3, were chosen (rather than other methods) was given in subsection 2.3.2.

minimizes the upper bound of the actual risk, compared with other classifiers that minimize the empirical risk."

Because Gordini and Velgio (2016) found favourable results for an RBF kernel, and the choosing of an SVM is largely based on their study, the current study will also make use of an RBF kernel to construct the hyperplane that separates churners and non-churners. Gordini and Velgio (2016) further explain that there are two relevant parameters to tweak for this specific kernel, being C and gamma. C is the penalty parameter for the error term and gamma is the kernel specific parameter. For this kernel, the gamma can be said to specify the 'range' of one training example. Higher values of gamma indicate a 'close' range and lower values indicate a 'far' range (Pedregosa et al., 2011). For a broad overview of the application of SVMs in a churning context the reader is referred to the works of Gordini and Veglio (2016), Kim and Sohn (2010), and Kim, Shin, and Park (2005).

3.5.3. Boosted Decision Tree

Boosting is a form of ensemble learning. The concept of ensemble learners was introduced by Breiman (1996). Breiman (1996) explains that by combining classifiers, model performance can increase significantly. There are several ensemble learners, but boosting seems to be preferred when it comes to churn analysis (Tamaddoni Jahromi et al., 2014; Burez & Van den Poel, 2009; Lemmens & Croux, 2006). Boosting operates by adding more weight to false positives and false negatives by running a base classifier (a decision tree in most cases) over several iterations. Eventually, those instances that are deemed 'hard' to classify become more important in the training process which leads to a more accurate model. Although several different version of boosting exist, the current study will only utilize adaptive boosting (hereafter AdaBoost). AdaBoost is chosen because it has proven successful in the study done by Tamaddoni Jahromi et al. (2014), on which the choosing of a boosted decision tree model is based (see section 2). For a more technical overview of how AdaBoost works, the reader is referred to the appendix of the work done by Tamaddoni Jahromi et al. (2014). For a relatively recent overview on the mechanics of simple decision trees, the reader is referred to a comprehensive overview written by Kotsiantis (2013).

To establish what parameters are of interest, the work of Ridgeway (2006) will be reviewed. This author proposes that, to properly construct an AdaBoost model with a decision tree as base learner, one should manipulate the number of base learners, the depth of the base learners, and the learning rate. Ridgeway (2006) explains that the number of base learners and the learning rate interact when it comes to the performance of the model. By decreasing the learning rate, the number of base learners required increases (because the learning rate is used to shrink the contribution of each tree as it is added to the model). In general, a large number of base learners with a small learning rate is seen as preferable, due to the relatively small 'step size' from one base learner to the next. The smaller the step size, the less information will be lost by iterating from one base learner to the next. Ridgeway (2006) goes on to explain that the learning rate also interacts with the depth of the base classifiers. By adding more nodes (more depth), fewer base learners are required to minimize error. So, as more depth is incorporated into the model, the learning rate should be further reduced for sufficient base learners to be fitted. Therefore, the number of base classifiers, depth of the base classifiers, and the learning rate will be tweaked to construct the best performing AdaBoost model.

3.6. Evaluation Techniques

To establish whether churn prediction is applicable to the current context, proper evaluation of the predictive analyses will be needed. Two evaluation methods will be employed. The first method makes use of the area under the receiving operating characteristic curve (hereafter ROC AUC score), which will be consulted to select the optimal parameters for each unique model/algorithm combination, based on 5-fold cross-validation on the training set. Next, all optimized model/algorithm combinations will be run on the test set, again producing an ROC AUC score for each algorithm/model combination. By doing so, a choice can be made between whether the standard model or the Bilderberg model is best suited to predict churn in the current context, based on which model produces the highest ROC AUC score on the test set. The second method will be to construct a cumulative gain chart for the optimized classifier of the chosen model, which will be used to assess the practical relevance of the optimal model.

3.6.1. ROC AUC

Due to the binary nature of the outcome variable, along with the uneven distribution of classes, The ROC AUC score is one of the most popular methods to evaluate the performance of classifiers in the churning context (Gordini & Veglio, 2016; Keramati et al., 2016; Jahromi et al., 2014). Burez and Van den Poel (2009) state that, aside from the fact that using the ROC AUC score is generally seen as the norm when evaluating churn prediction models, ROC AUC scores are especially useful because AUC does not place more emphasis on one class over the other. This means that the ROC AUC score, unlike other evaluation techniques, is not biased against the minority class. To inform the reader of what the ROC AUC score is and what it does, a brief overview of the work done by Fawcett (2006) will be offered. In order to get a more technical and comprehensive understanding of this evaluation metric, the reader is referred to the full work of this author.

Fawcett (2006) explains that the ROC curve is a two-dimensional graph in which the true positive rate is plotted against the false positive rate. Therefore, the ROC curve is said to be visualizing a trade-off between benefits (true positives) and costs (false positives). Basically, in the case of binary classification, the more northwest the line representing a classifier is on the plot relative to the diagonal baseline, the better the classifier is said to be discriminating between the two categories. The difference between the lines of the classifier and the baseline can be quantified by measuring the area under the curve (AUC) of the line representing the classifier. As the AUC is a portion of the area of the unit square, it will always be between 0 and 1. However, as the baseline always has a value of 0.5, it is expected that the value of a valid classifier ranges between 0.5 and 1. So the higher the ROC AUC score, the better the classifier is discriminating between the two classes. Also, as the baseline of the results of a random classifier on a binary problem is 0.5, an algorithm is only considered to be valid when its ROC AUC score is considerably higher than this baseline.

As mentioned above, ROC AUC scores will be used to find the best parameters for each model/algorithm combination by cross-validating (5-fold) on the training set. Subsequently, all model/algorithm combinations will be run on the test set, ultimately producing 6 ROC AUC scores that are considered valid, and thus applicable to churn analysis in the non-contractual B2B hotel industry, if the ROC AUC score is at least higher than the 0.5 baseline. Based on these 6 ROC AUC scores, the model that is most applicable to the current context can also be chosen. This would be the model/algorithm combination that produces the best ROC AUC score. As was explained in subsection 3.2.2, the Bilderberg model will only be considered best if the ROC AUC score of its optimal algorithm is *significantly* higher than the ROC AUC score of the optimal algorithm of the standard model. If the performance of the Bilderberg is worse, the same, or only insignificantly higher than the standard model, the standard model will be considered best. The reader is reminded that this is the case because the Bilderberg model is an expansion of the standard model (see section 3.2).

3.6.1. Cumulative Gain Chart

In order to see how the optimal model does in practice, a cumulative gain chart will be constructed on the basis of its performance on the test set. The cumulative gain chart is popular throughout marketing literature with a practical focus, due to its straightforwardness when it comes to managerial interpretation (Jahromi et al., 2014). In this context, the cumulative gain chart can be graphed by obtaining the posterior classification probabilities of each company as these were obtained from the test set on which the predictive algorithms were ran. These probabilities are then ordered from highest to lowest. Next, the true amount of churners are counted separately for ten decimals of the ordered data. Finally, the percentage of true churners identified can be plotted against the percentage of companies that need to be targeted in, for example, a retention campaign. In this graph, just like in the ROC AUC graph, a diagonal baseline can be identified that, with a random classifier, identifies the same percentage of true churners as the percentage of companies targeted in a retention campaign. Plotted curves of properly functioning algorithms are expected to be more northwest of this baseline, indicating the identification of more true churners with less companies targeted in a retention campaign. By implementing the model, Bilderberg could target more churners with less resources, and thus reduce costs (Jahromi et al., 2014).

3.7. Descriptive Analysis

In order to rank features from most important to least important, sklearn offers several techniques that differ per predictive algorithm. The ranking of features is required to answer the second research question, which asks what features are most important to predict churn in the current context. These different techniques will be described per predictive algorithm used. Note that these techniques will be applied to the training set with optimal parameters for each model/algorithm combination, as they were obtained through cross-validation. If these techniques appear to bear similar results per model, findings on feature importance for churn analysis in the hotel industry can be said to be robust. Below, each descriptive method (differing per predictive algorithm) is further elaborated on.

The data has been standardized to Z-scores (see above), and LR falls under the category of linear models. This means that the coefficients of the features used in constructing these models can be seen as a proxy of feature importance, with higher coefficients indicating more importance. Furthermore, if a coefficient is above 0, it indicates a positive effect. Consequently, if a coefficient is below 0, it indicates a negative effect. As the methods described below do not indicate a direction of the effect, the results of the LR coefficient analysis will be seen as decisive for the direction of the effect (Pedregosa et al., 2011).

For an SVM with a non-linear kernel, there are no straightforward ways to hierarchically rank features (Maldonado, Weber, & Basak, 2011). Therefore, ablation (as was explained in section 2) will be used to assess feature importance for the SVM. However, instead of an increasing error rate, the current study will make use of declining ROC AUC score rates, with the highest rates resembling the most important features. This choice was made because the ROC AUC measure is the most important evaluation technique used in the current investigation.

AdaBoost can rank features hierarchically by assigning them a value between 0 and 100. All these values together sum to 100, indicating that the higher the value of the feature, the more important the feature is.

This process is based on how often (on average) each feature is used in the split points of a weak classifier (Pedregosa et al., 2011).

3.8. Answering Research Questions and Addressing Problem Statement

In the next paragraph, answers (in the form of results) will be given to the research questions. The first research question asked whether extending a standard churn prediction model to the Bilderberg model would be an improvement when predicting churn in the non-contractual B2B hotel industry. This research question will be answered by comparing ROC AUC scores of the standard and Bilderberg models for all algorithms (as they were obtained from running all optimized model/algorithm combinations on the test set). If the Bilderberg model, as it was analysed with the best predictive algorithm, does not obtain a *significantly* higher ROC AUC score than the standard model, this is not the case. The second research question asked what predictors are most important when it comes to predicting churn in the non-contractual B2B hotel industry. This question will be answered by comparing feature importance for both models (standard and Bilderberg) across all three predictive algorithms (LR, SVM, and AdaBoost). The problem statement will be addressed briefly in the next section by comparing the ROC AUC score of the optimal model/algorithm combination (as it was obtained from the test set) with the 0.5 baseline, as well as elaborating on the results of the cumulative gain chart. A more in-depth analysis with regard to the problem statement will be offered in section 5.

4. Results

In this section, the results of the analyses described in section 3 will be discussed. The goal of this section is to empirically explore the possibilities of constructing a churn prediction model in the non-contractual B2B hotel industry. First, in subsection 4.1, the results of all three algorithms for each model (standard and Bilderberg) will be optimized on the training set[9]. Subsequently, the performance of all the optimized model/algorithm combinations on the test set can be compared in order to answer the first research question. Second, in subsection 4.2, feature importance will be analysed for both models (standard and Bilderberg), by utilizing the descriptive techniques that were described in subsection 3.7. This will produce 6 lists of feature importance (2 models analysed with 3 descriptive methods each) and an overview of the directions of the effects that can be used to answer the second research question. Finally, in subsection 4.3, the applicability of the optimal model/algorithm combination is assessed. This will be achieved by comparing the ROC AUC score of the best scoring algorithm of the optimal model (as it was obtained from running this model/algorithm combination on the test set) with the ROC AUC score baseline (which was determined in subsection 3.6). Furthermore, a cumulative gain chart will be constructed to see how the churn prediction analysis does in 'practice'. Conclusions with regard to the problem statement will be based on these findings and will be discussed in section 5.

4.1. Comparing Models (RQ 1)

The goal of this subsection is to provide an answer to RQ 1, which asked whether a previously proven successful (in different contexts) and generally applicable standard model needed to be extended to a Bilderberg model (with more context-specific features) in order to analyse churn in the non-contractual B2B hotel industry. This subsection will begin with a presentation of the results of parameter optimization on the training set for both models of the logistic regression (subsection 4.1.1), followed by the results of parameter optimization for both models of the support vector machine (subsection 4.1.2), and finally the results of parameter optimization for both models of the boosted decision tree (subsection 4.1.3). As was explained section 3, this optimization process will be based on cross-validated (5-fold) ROC AUC scores. Naturally, the parameter settings that produce the highest ROC AUC score per model/algorithm combination will be considered best. Next, choosing which algorithm does best per model, and subsequently picking the best model, will be based on additional ROC AUC scores as they were obtained from running all 6 optimized model/algorithm combinations on the test set (subsection

[9] In order to determine under what parameters these classifiers performed best, parameter optimization was done by using a grid search package (sklearn) with 5-fold cross-validation. The eventual choice of parameters was, just like the eventual choice of algorithms and models, based on the settings that produced the highest ROC AUC score.

4.1.4, RQ 1). The standard model is considered to be best when its ROC AUC score (on the test set) is about the same or higher than the ROC AUC score of the Bilderberg model (on the test set). The Bilderberg model is considered to be best when its ROC AUC score (on the test set) is *significantly* higher than the ROC AUC score (on the test set) of the standard model.

4.1.1. Results of Parameter Optimization: Logistic Regression

As was explained in section 3, the level of C is the relevant parameter to tune for an LR. The LR was performed on the training data (with 5-fold cross-validation) for levels of C that increased exponentially ranging from 0.0001 to 10. Performing a grid search indicated that a C value of 0.01 yielded the best result for both models (standard and Bilderberg). Several more specific grid searches for the standard model indicated an optimal C value of 0.0091 with an ROC AUC score of 0.8080, and several more specific grid searches for the Bilderberg model indicated an optimal C value of 0.0096 with an ROC AUC score of 0.8051.

4.1.2. Results of Parameter Optimization: Support Vector Machine

As stated in section 3, relevant parameters to tune for an SVM with an RBF kernel are C and gamma. An SVM was performed on the training data (with 5-fold cross-validation) for levels of C and gamma that increased exponentially from 0.0001 to 100. A grid search indicated the best results with a C value of 1 and a gamma value of 0.01 for the standard model, and a C value of 100 and a gamma value of 0.0001 for the Bilderberg model. Additional specific grid searches for the standard model yielded the best results with a C value of 1.3 and a gamma value of 0.013, and for the Bilderberg model a C value of 90 and a gamma value of 0.000095. With these parameters, the SVM produced ROC AUC scores of 0.8068 and 0.8055 respectively for the standard and Bilderberg models.

4.1.3. Results of Parameter Optimization: Boosted Decision Tree

As stated in section 3, optimal performance of AdaBoost can be accomplished by altering the learning, rate, the number of base learners, and the depth of the base learners. An AdaBoost was performed on the training data (with 5-fold cross-validation) for levels of learning rate that increased exponentially between 0.0001 and 10, numbers of base learners that ranged between 100 and 2000 (increasing with steps of 100), and a depth of the base learners that ranged between 1 and 10 (increasing with steps of 1). Grid search indicated optimal performance with a learning rate of 0.01, 1200 base learners, and a depth of 1 for the standard model. The Bilderberg model obtained optimal performance with a learning rate of 0.01, a depth of 3, and 100 base learners. Finer tuned grid searches for the standard model yielded an ROC AUC score of 0.8022 when parameters were set to 1243 base learners, a learning rate of 0.006, and a depth of 1. Finer tuned grid searches for the Bilderberg model resulted in an ROC AUC of 0.8017 under a parameterization of 6900 base learners, a learning rate of 0.001, and a depth of 1.

4.1.4. Summary of Experimental Findings: RQ 1

With the parameters optimized for all model/algorithm combinations, an analysis of the model/algorithm combinations on the test set was possible. A summary of the results that were obtained from this analysis is depicted in table 1. The LR appears to be the best performing algorithm for the standard model. However, it should be noted that the results of the AdaBoost algorithm are not far behind. The results of the SVM are disappointing compared to the results of cross-validation, and substantially lower than the results of the other algorithms. For the Bilderberg model, the LR appears to do best, again closely followed by the AdaBoost. And, again, the SVM displays disappointing results, however not as bad as it did for the standard model. A possible explanation for the disappointing results of the SVM will be offered in section 5. Another interesting find is the fact that the results for both the standard- and the Bilderberg model are the same for the AdaBoost algorithm. An explanation for this phenomenon will be offered in the next subsection on feature importance.

Table 1

Algorithm Performance per Model Measured in ROC AUC Score

	ROC AUC	
Algorithm	Standard Model	Bilderberg Model
LR	**0.7464**	0.7442
SVM	0.5042	0.6054
AdaBoost	0.7416	0.7416

Note. ROC AUC scores were established by running all optimized model/algorithm combinations on the test set. All scores are above the 0.5 baseline, however the SVM that was used on the standard model only barely scored above the baseline.

So, for both models the LR appears to yield the best results, producing figures that are not far apart. Still, the conclusion can be drawn that the standard model performs best, and therefore an expansion of it is not needed to analyse churn in the current context. This conclusion is drawn on the ground of several arguments. First, the absolute ROC AUC score on the test set of the standard model is simply higher than that of the Bilderberg model. Second, the standard model can be generalized much easier to any business operating in the hotel industry that has access to a transactional (CRM) database. Third, due to the lower amount of predictor variables, the process of isolating and creating features, as well as the process of parameter optimization is much more time efficient.

4.2. Feature Analysis (RQ 2)

In this subsection, an overview of feature importance for churn analysis in the non-contractual B2B hotel industry will be given. By doing so, an answer to RQ 2 (what features are most important for analysing churn in the current context?) can be given. As was explained in section 3, for each algorithm a different method to hierarchically order features was used. These methods were looking at the standardized coefficients (LR), ablation (SVM), and the number of average split points per base learner scaled from 1 to 100 (AdaBoost). As was also explained in section 3, only the standardized coefficients can be used to indicate the direction of the effects (but only for effects that are significant). The indication of the direction of the effects produced by the standardized coefficients is considered to be adequately guiding because the LR performed best for both models. Therefore, an overview of the descriptive findings of this algorithm is offered first. A subsequent overview of, respectively, the SVM and AdaBoost algorithms will be offered to establish whether the results of the feature analysis concur.

4.2.1. Descriptive Results: Logistic Regression

In table 2, the results of the descriptive analysis on the standard model with the LR algorithm are presented. As can be seen, ‘recency’, by far, is the most important predictor variable in this analysis. Furthermore, the other two features related to a point in time, ‘observation_period’ and ‘interpurchase’, are the second and third most important features respectively, indicating the importance of the time dimension in general.

Table 2

Results of descriptive analysis: LR, standard model.

Features (ranked)	Standardized Coefficients
recency	0.838
observation_period	-0.212
interpurchase	0.176
ors_revenue	-0.155
frequency	-0.142

Note. Features are ranked hierarchically from most important to least important.

In table 3 (depicted on the next page), the results of the descriptive analysis on the Bilderberg model with the LR algorithm are presented. As can be seen, ‘recency’, again, plays the most important role by far. An interesting finding is that the ‘nr_locations’ variable (how many different Bilderberg locations the company visited) seems to be more important than the other (than ‘recency’) variables related to time, ‘observation_period’ and ‘interpurchase’, which respectively obtained a second and third place of importance when the LR was used to descriptively analyse the standard model (although these features still obtain third and fourth place in the Bilderberg model).

Table 3

Results of descriptive analysis: LR, Bilderberg model.

Features (ranked)	Standardized Coeffcients
recency	0.794
nr_locations	-0.191
observation_period	-0.182
interpurchase	0.150
frequency	0.124
ors_nr_of_adults	0.100
booker	0.087
ors_revenue	-0.081
ors_nr_of_rooms	0.049
ors_length_of_stay	0.022
Type	0.003

Note. Features are ranked hierarchically from most important to least important.

As was explained in section 3, by analysing the results of the LR further, the direction of the effects can also be assessed. P-values[10] were evaluated on the basis of an alpha of .050. For the standard model, only the variables related to time displayed significant effects. For the Bilderberg model, again, variables related to time had significant effects, but the 'nr_of_locations' variable appeared to have a significant effect as well. Therefore the directions of these effects will be discussed next. 'Recency' has a positive effect ($p < .001$), meaning that the higher the recency (thus, the further away the last hotel visit was), the higher the chance to churn becomes. This is in line with theoretical and statistical findings presented in section 2 and 3. Similarly, the length of the observation period of a company has a negative effect ($p < .001$). This means that the longer the observation period (thus, the longer the amount of time since the first booking in the calibration period), the less likely the chance for a company to churn becomes. The 'interpurchase' feature has a positive effect ($p < .001$). This means that if the period between purchases of a company becomes larger, the chance to churn increases as well. This finding is also in line with theoretical findings presented in previous sections. Finally, the 'nr_of_locations' variable had a negative effect ($p = 0.02$), indicating that the more Bilderberg locations were visited by a company, the less likely the company would churn. Although no theoretical basis was found for adding the 'nr_of_locations'

[10] P-values were obtained by using the Stats-Models Logit API, as sklearn does not include significance testing in their LR API.

variable, the statistical findings presented in section 3 appear to be in line with the finding that it has a significant negative effect.

4.2.2. Descriptive Results: Support Vector Machine

In table 4, the results of the descriptive analysis on the standard model with the SVM algorithm are presented. As can be seen, the results (to a certain extent) mirror the results of the LR algorithm on the standard model. Again, 'recency' is the most important variable, with a relatively small role for other time-related variables. Monetary value and frequency barely seem to play a role.

Table 4

Results of descriptive analysis: SVM, standard model.

Features (ranked)	Reduction in ROC AUC
recency	0.085
observation_period	0.004
interpurchase	0.002
ors_revenue	0.000
frequency	0.000

Note. Features are ranked hierarchically from most important to least important.

The results of the descriptive analysis on the Bilderberg model with the SVM algorithm are given in table 5 (depicted on the next page). Again, 'recency' is very important, with a relatively small role for the other time related variable. A remarkable finding is that the 'nr_locations' variable does not seem to play a role in the SVM, whereas it did in the LR. Furthermore, all variables that are not related to time do not seem to play a role.

Table 5

Results of descriptive analysis: SVM, Bilderberg model.

Features (ranked)	Reduction in ROC AUC
recency	0.084
observation_period	0.003
interpurchase	0.002
ors_nr_of_adults	0.000
frequency	0.000
booker	0.000
ors_length_of_stay	0.000
nr_locations	0.000
ors_nr_of_rooms	0.000
Type	0.000
ors_revenue	0.000

Note. Features are ranked hierarchically from most important to least important.

4.2.3. Descriptive Results: Boosted Decision Tree

In table 6, the results of the descriptive analysis of the AdaBoost algorithm on the standard model are given. As can be seen, ‘recency’, again, is by far the most important variable. An interesting difference with the LR and SVM is that the other time related features (‘observation_period’ and ‘interpurchase’) are not next in line. In fact, these features, just like monetary value, appear to be useless in this analysis. Instead, frequency is second highest when it comes to feature importance.

Table 6

Results of descriptive analysis: AdaBoost, standard model.

Features (ranked)	Relative use in splitting points.
recency	87.0%
frequency	13.0%
observation_period	0.0%
interpurchase	0.0%
ors_revenue	0.0%

Note. Features are ranked hierarchically from most important to least important.

The descriptive results of the AdaBoost algorithm as it was run on the Bilderberg model are presented in table 7 (depicted on the next page). Once again, ‘recency’ is by far the most important variable. Moreover, it appears that ‘nr_locations’, although it was the second most important variable in the LR, has no effect, and ‘frequency’ is, again, next in line. This appears to be comparable to finding no effects for the ‘observation_period’ and ‘interpurchase’ variables in the standard model, which did play an important role in the LR. In section 5, a possible explanation will be given as to why the findings between the LR and SVM algorithms on the one hand, and the AdaBoost algorithm on the other might differ. As stated in section 4.1.4, the ROC AUC score of both AdaBoost models (standard and Bilderberg) were excactly the same. This can probably be explained by the fact that the same amount of value was attached to ‘recency’ and ‘frequency’ in both models, with no value being attached to the other predictors in both models at all. This indicates that expanding the standard model literally has no effect when the AdaBoost algorithm is used.

Table 7

Results of descriptive analysis: AdaBoost, Bilderberg model.

Features (ranked)	Relative use in splitting points.
recency	87.0%
frequency	13.0%
interpurchase	0.0%
ors_nr_of_adults	0.0%
observation_period	0.0%
booker	0.0%
ors_length_of_stay	0.0%
nr_locations	0.0%
ors_nr_of_rooms	0.0%
Type	0.0%
ors_revenue	0.0%

Note. Features are ranked hierarchically from most important to least important.

4.2.4. Summary of Experimental Findings: RQ 2

All descriptive analyses have in common that ‘recency’ is by far the most important feature when it comes to predicting churn (for both models). Furthermore, other time elements (the length of the observation period and the time between purchases) appear to be next in line for the LR and the SVM algorithms in the standard model, whereas the Bilderberg model displays the same feature importance for the SVM but not for the LR (with ‘nr_locations’ being next in line after ‘recency’ here). More importantly however, the AdaBoost algorithm favours frequency of purchases over the other time elements (standard model) and ‘nr_locations’ (Bilderberg model). With regard to the research question that asks which features are most important for analysing churn in the non-contractual B2B hotel industry, the following results are most relevant. First, ‘recency’ is very important for analysing churn in this context. Therefore, this feature should be incorporated in any model. Second, other time related features (‘observation_period’ and ‘interpurchase’), ‘nr_locations’, and ‘frequency’ can be of importance depending on the model/algorithm combinations that are being used. Third, monetary value and context-specific variables other than ‘nr_locations’ seem to play the least important roles when churn is analysed in the current context. A further elaboration on these findings will be offered in section 5.

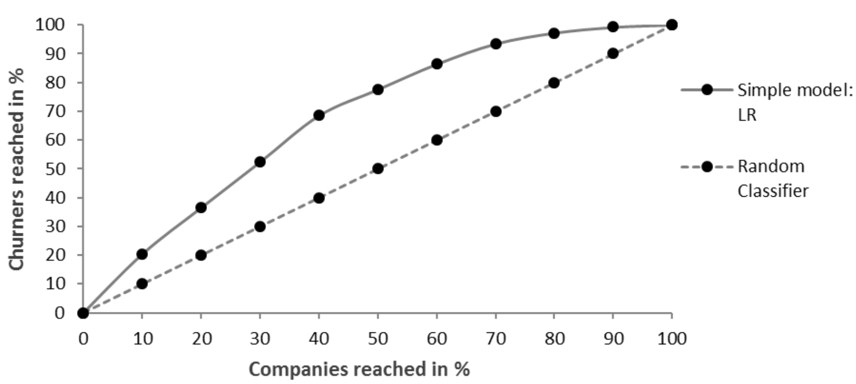
4.3. Applicability of Optimal Model

In this subsection, the performance of the optimal model (standard, analysed with the LR algorithm) is further evaluated. As stated in subsection 4.1.4, predicting churn with this model resulted in an ROC AUC score of 0.7464 on the test set. This score is substantially higher than the 0.5 baseline, and thus valid.

The cumulative gain chart constructed on the basis of the standard LR model (here called Simple Model: LR) is depicted in figure 9. As can be seen, the model has practical relevance. If Bilderberg were to target 50% of their B2B customer base with a retention campaign, they would reach about 80% of the churners by using this model. If they would not use the standard LR model, but would continue to prevent churn at random instead, they would only reach 50% of the churners when they target 50% of their B2B customer base.

Figure 9

Cumulative gain chart



5. Discussion

In this section, the results of the current study will be evaluated. First, in subsection 5.1, the research questions as they were constructed in section 1 will be fully answered. Second, in subsection 5.2, the problem statement (as it was constructed in section 1) will be addressed. Addressing the problem statement will be based on the results of the optimal predictive analysis on the test set, in combination with conclusions drawn from the research questions. Third, in subsection 5.3, the current study is placed in context by elaborating on the contributions it has made. Finally, in subsection 5.4, recommendations for future research will be made.

5.1. Answering Research Questions

The current study explored possibilities to analyse churn in the hotel industry, and, indirectly, the possibilities of data-driven CRM in this industry in general. Researching this subject was highly needed due to a lack of it in the current hotel industry literature, even though it has long been known that not adequately dealing with churn can significantly impair a company's business perspectives (Reichfeld & Sasser, 1990), a shift in importance from customer acquisition to customer retention is apparent in every industry (Hadden et al., 2007), and constructing a proper churn prediction model might increase the average net present value of customers (Tamaddoni Jahromi et al., 2010).

To narrow down the general notion of churn analysis in the hotel industry, the context was further specified to being non-contractual and B2B. Both types of churn (non-contractual and B2B) have been investigated before in other industries and researchers have been able to find good results (for example Jahromi et al., 2014). The non-contractual restriction was imposed because most of Bilderberg's customer base consists of customers who have a non-contractual relationship with the hotel. The B2B restriction was imposed because it was speculated that B2B customers generally have a more long term and loyal relationship with the hotel (Claycomb et al., 2005). To further concretize the current research, two research questions were constructed. These research questions were:

Research question 1: *Is the extension of a standard churn prediction model by incorporating context-specific features useful when analysing churn in the non-contractual B2B hotel industry?*

Research question 2: *Which predictors are considered most important in order to predict churn in the non-contractual B2B hotel industry?*

These research questions will be answered in the next two subsections.

5.1.1. Answering Research Question 1

To investigate whether churn can be predicted adequately with a model that easily generalizes to other contexts, keeps computing times to a minimum by including as few variables as possible, and had been proven successful in different contexts, the standard model was created. The standard model was tested against a model that contained extra variables that were more specific for Bilderberg (the Bilderberg model). In this sense, the Bilderberg model could be seen as an extension of the standard model (coining the standard model as the baseline). The extension of the standard model was made due to favourable findings of an exploratory data analysis with regard to several context-specific features (see section 3). Because the standard model has several advantages over the Bilderberg model (such as computing time and generalizability) the Bilderberg model would have to perform *significantly* better than the standard model (in terms of ROC AUC score) for it to be considered superior.

The performances of the standard- and Bilderberg models were compared by running three predictive algorithms (LR, SVM, and AdaBoost) on the same dataset. Parameters were tuned for each algorithm to specifically match one of the two models (5-fold cross-validation on the training set). Evaluation was based on which model/algorithm combination produced the highest ROC AUC score on the test set. It appeared that the LR produces a better ROC AUC score for the standard model, also obtaining the highest overall score. The SVM did better with the Bilderberg model, however, the test set scores of this algorithm for both models were substantially lower than the scores of the two other algorithms. A possible explanation for this phenomenon might be that running an SVM, especially with an RBF kernel, might be too much of an overkill (the SVM might have been overly fit to the training set), seeing as a simple LR performs better. In this sense, the composition of the data appeared to fit best to a relatively simple model. This might also explain why the SVM did better with the more complex Bilderberg model and the other algorithms did not. Finally, the AdaBoost algorithm obtained the same ROC AUC score for both types of models. This indicates that expanding the standard model with the chosen context-specific predictor variables literally has no effect when AdaBoost is used.

All in all, it can be concluded that these findings mean that an extension of the standard model with more context-specific predictor variables did not appear to be necessary to analyse churn in the non-contractual B2B hotel industry. This finding is in accordance with the work done by Jahromi et al. (2014), who, in order to predict churn in a non-contractual context, suggest the use of a limited set of predictors with maximum predictive power that are easily generalized to other databases.

5.1.2. Answering Research Question 2

Model performance was further explored by analysing which features were of the highest importance when predicting churn in the current context. Furthermore, because an LR was used, the direction of the effects of each feature could also be determined based on the coefficients that were found. The directions of the effects, as they were established by the LR, are seen to be especially guiding because the performance of this algorithm appeared to be best for both models. The direction of the effects will be discussed next, after which an evaluation of feature importance will be given.

For the standard model, most findings on the direction of the effects of the features are in line with theory and expectations formulated in section 2. The less recent a customer made a booking, the higher the chance of churning became (as was also found by Coussement & De Bock, 2013). As expected, the opposite appeared to be true for the length of the observation period, with higher values indicating a lower chance to churn. Furthermore, as was also found by Buckinx and Van de Poel (2005), less time between purchases resulted in a lower propensity to churn. Note that these confirmed findings all have in common that the variables are related to a point a time. Findings on frequency of purchases and revenue generation per company were in line with theory and expectations formulated in section 2, in the sense that the effects were in the direction that were expected (meaning that both more ‘frequency’ and ‘ors_revenue’ pointed to a reduced chance to churn). However, these effects cannot be generalized, as they did not appear to be significant. For the Bilderberg model, again, variables related to a point in time appeared to be significant, and displayed the expected effects. An interesting finding was that the variable that indicated the amount of different Bilderberg locations that were visited by a company also explained a significant part of the variance. It was established that the more locations were visited by a company, the less likely it would be for the company to churn. Not finding significant effects for the other context-specific variables is not surprising, seeing as the performance of the Bilderberg model was generally worse than the performance of the standard model.

In order to assess the importance of each individual feature, features were ranked hierarchically for all three predictive algorithms. As was mentioned in section 4, ‘recency’ of the latest purchase is the most important feature for all the model/algorithm combinations. Other time related features (‘observation_period’ and ‘interpurchase’) appear to be next in line for both the LR and SVM (note that ‘nr_locations’ was next in line for the LR that was utilized in combination with the Bilderberg model), whereas the AdaBoost model favours the frequency of purchases (for both models). Differences, between the LR and SVM on the one hand and the AdaBoost on the other, can possibly be explained by the fact that LR and SVM are very similar algorithms in general (Salazar, Velez, & Salazar, 2012), whereas

AdaBoost specifically focuses on hard-to-classify examples (Ridgeway, 2006). In this sense, an explanation might be that low levels of ‘frequency’ tended to vary between the classes, more than did, for example, low levels of ‘recency’. This can be seen as an indication that examples with extreme values of ‘frequency’ are probably perceived as hard-to-classify examples, and thus are weighted heavier by the AdaBoost algorithm.

A general conclusion that can be drawn is that transactional features related to time should always be included in an analysis in the current setting (especially the ‘recency’ feature), and that features related to frequency should be included when any form of boosting algorithm is being used. Furthermore, when following the suggestions made by Jahromi et al. (2014) to construct a churn prediction model for the current context (whose study served as the basis to construct the standard model), monetary value could be excluded from the model. Also, when constructing a more context-specific model for the current context, if possible, a feature related to how many different hotel locations were visited by a company should be included.

5.2. Addressing the Problem Statement

By answering the research questions above (and further evaluating them below), the problem statement of the thesis can be addressed. The problem statement was formulated in section 1 and entailed the following:

Problem Statement: *To what extent is the construction of a churn prediction model viable in the non-contractual B2B hotel industry?*

In the course of answering the research questions, all model and algorithm combinations were compared on the basis of their ROC AUC scores. This process indicated that the standard model in combination with the LR algorithm was optimal, producing an ROC AUC score on the test set that was substantially higher than the baseline, indicating a valid model. Furthermore, from a managerial perspective, the cumulative gain chart indicated that 80% of churners could be reached by targeting 50% of the companies in the database, clearly indicating practical usefulness of the model.

As the standard model was considered best (RQ 1), and generally consists of transactional features, the model can easily be applied to any hotel business that maintains a transactional database with timestamps. This also means that, when analysing churn, the hotel industry probably is not very different from other industries in which churn analysis has been viable such as E-Commerce (Jahromi et al., 2014). Furthermore, by analysing feature importance in depth (RQ 2), it appeared that especially the ‘recency’ predictor variable is important when distinguishing potential churners from potential non-churners in the current context. By examining the directions of the effects, a profile of a potential churner could also be constructed. In this sense, a potential churner could be identified as being a company with a purchase history that is not very recent, that has a short period of observation, and displays a relatively high time between purchases. It should however be noted that, if a more context-specific model is applied, the profile of a churner can be extended to being a company that, other than the specifications mentioned above, visited only one or a few Bilderberg locations.

As the goal of analysing churn is to apply the best functioning model to anticipate churn, as well as investigate what factors are most indicative of churn, the current study is seen as having been successful in the non-contractual B2B hotel industry. Therefore the definitive answer to the problem statement is: yes, churn analysis is viable in the non-contractual B2B hotel industry, to the extent that an optimal model (being a standard model in combination with the LR algorithm) outperforms the baseline, yields practical relevance, and gives a clear indication of what predictor variables are important.

5.3. Contributions of Study

The current study explored the possibilities of predicting churn in a context that had not been investigated before (the non-contractual B2B hotel industry), and was meant to decrease the general lack of data-driven CRM research in this context. By successfully constructing a churn prediction model, it has been shown that churn can indeed be analysed in the current context. As was pointed out above, the construction of the best performing model was based on including only a few simple predictors (with a previously proven high amount of predictive power) that are not bound to the case under investigation specifically. Therefore, findings of the current study are definitely generalizable to other hotels, and possibly even to other types of industries that have not been investigated before, as long as access to a transactional database can be granted.

From a practical point of view, Bilderberg now has the tools to connect with its customers in a more efficient manner. Instead of going with a gut feeling, potential churners can now be contacted before actual churn has arisen. If these customers are persuaded to maintain the relationship with the hotel, a positive effect on revenue might become apparent (Tammaddoni Jahromi et al, 2010). Bilderberg now also knows which factors are of the highest importance when signalling churn in time is crucial. Finally, Bilderberg has a better of understanding of how they should maintain their database in order for data-driven CRM research to be possible.

5.4. Directions for Future Research

Based on the process of constructing a churn prediction model in the non-contractual B2B hotel industry, several recommendations for further research can be done. Firstly, it is suggested that a similar analysis is conducted, but on a database with a larger time window. As was discussed in section 3, the ideal time window was larger than was possible due to the lack of further historical data in the database. It is expected that, with a larger time window, results would improve, meaning that churning can be identified in advance even more accurately.

Secondly, the choice to solely investigate the B2B market was made because it was expected that customer loyalty would be higher, and therefore a churn analysis would be more valid. It would be interesting to see whether this really is the case in a study that compares B2B and B2C in the non-contractual hotel industry. It would also be interesting to see whether methods used in the current study can be generalized to the B2C context.

Thirdly, an additional study could be conducted with features extracted from an external database (instead of purely focussing on an internal database as this thesis has done) that could possibly extend the standard model even further. For example, by coupling a general business database, one might be able to add features such as the business context the companies are in. By doing so, an extended version of the standard model may be preferred after all.

Fourthly, ideally, each vector should belong to one unique company. In this sense, bias is avoided as much as possible. This was attempted in the thesis, but because the registration of companies to the database was not bound to a formal process, ideal identification could not be achieved. In subsequent research it is suggested to work with a database that was built with a system that automatically couples companies (or unique individual customers) to each unique vector.

Fifthly, employing additional evaluation metrics such as accuracy might proof to be fruitful in future research. For example, if an even closer collaboration with the case under investigation can be reached, it could be interesting to investigate the exact impact of false positives and false negatives on revenue. To establish this impact, more data related to the costs of retention campaigns in practice is needed, in combination with the exact revenue boost of identifying a churner. Ideally, the established revenue boost per identified churner could be personalized further by establishing (clustered) rates of customer value (for an example of doing this for a recommender system in banking see: Schmitt, Skiera, & Van den Bulte, 2011). Ultimately, weights of the predictive algorithms could hypothetically be altered to attach more value to false positives or negatives depending on the net present value of a churner (Tamaddoni Jahromi et al., 2010), as it was established on the basis of the factors mentioned above.

References

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications policy, 30*, 552-568.

Akinci, S., Kaynak, E., Atilgan, E., & Aksoy, Ş. (2007). Where does the logistic regression analysis stand in marketing literature? A comparison of the Market Positioning of Prominent Marketing Journals. *European Journal of Marketing, 41*, 537-567.

Anil Kumar, D., & Ravi, V. (2008). Predicting credit card customer churn in banks using data mining. *International Journal of Data Analysis Techniques and Strategies, 1*, 4-28.

Ben-Hur, A., & Weston, J. (2010). A user’s guide to support vector machines. *Data mining techniques for the life sciences*, 223-239.

Bolton, R. N., Kannan, P. K., & Bramlett, M. D. (2000). Implications of loyalty program membership and service experiences for customer retention and value. *Journal of the academy of marketing science*, *28*, 95-108.

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research, 164*, 252-268.

Bucklin, R. E., & Gupta, S. (1992). Brand choice, purchase incidence, and segmentation: An integrated approach. *Journal of Marketing Research*, *29*, 201.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, *36*, 4626-4636.

Caruana, A. (2002). Service loyalty: The effects of service quality and the mediating role of customer satisfaction. *European journal of marketing, 36*, 811-828.

Claycomb, C., Iyer, K., & Germain, R. (2005). Predicting the level of B2B e-commerce in industrial organizations. *Industrial Marketing Management*, 34, 221-234.

Duan, W., Cao, Q., Yu, Y., & Levy, S. (2013). Mining online user-generated content: using sentiment analysis technique to study hotel service quality. In *System Sciences (HICSS), 2013 46th Hawaii International Conference*, 3119-3128.

Fathian, M., Hoseinpoor, Y., & Minaei-Bidgoli, B. (2016). Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods. *Kybernetes, 45*, 732-743.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*, 861-874.

Gordini, N., & Veglio, V. (2016). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management*, *26*, 201-209.

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research, 34,* 2902-2917.

Hopmann, J., & Thede, A. (2005). Applicability of customer churn forecasts in a non-contractual setting. *Innovations in classification, data science, and information systems*, 330-337.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Jahromi, A. T., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management, 43*, 1258-1268.

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation, 2,* 10.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, 1-23.

Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining.* John Wiley & Sons.

Magnini, V. P., Honeycutt, E. D., & Hodge, S. K. (2003). Data mining for hotel firms: Use and limitations. *The Cornell Hotel and Restaurant Administration Quarterly, 44,* 94-105.

Maldonado, S., Weber, R., & Basak, J. (2011). Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences, 181,* 115-128.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*, 417-473.

Meyer-Waarden, L., & Benavent, C. (2009). Grocery retail loyalty program effects: self-selection or purchase behavior change? *Journal of the Academy of Marketing Science, 37*, 345-358.

Miguéis, V. L., Van den Poel, D., Camanho, A. S., & e Cunha, J. F. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert systems with applications, 39,* 11250-11256.

Min, H., Min, H., & Emam, A. (2002). A data mining approach to developing the profiles of hotel customers. *International Journal of Contemporary Hospitality Management, 14*, 274-285.

Mohammed, A. A., & Rashid, B. (2012). Customer Relationship Management (CRM) in Hotel Industry: A framework proposal on the relationship among CRM dimensions, Marketing Capabilities, and Hotel performance. *International Review of Management and Marketing, 2*, 220.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on neural networks, 11*, 690-696.

Murphy, D. (2007). Marketing for B2B vs. B2C – Similar but Different. Retrieved from: https://masterful-marketing.com/marketing-b2b-vs-b2c/

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research, 43,* 204-211.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications, 36*, 2592-2602.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Popović, D., & Bašić, B. D. (2009). Churn prediction model in retail banking using fuzzy C-means algorithm*. Informatica*, *33*, 235-239.

Pudipeddi, J. S., Akoglu, L., & Tong, H. (2014). User churn in focused question answering sites: characterizations and prediction. *Proceedings of the 23rd International Conference on World Wide Web* (pp. 469-474).

Reichheld, F. and Sasser, W. 1990. Zero defection: Quality comes to services. *Harvard Business Review, 68*: 105–111.

Reijden, van der, D. (2016, May, 23th). Horeca Top 100 2016 nummer 23: Bilderberg Groep. *Misset Horeca.* Retrieved from: http://www.missethoreca.nl/horeca/artikel/2016/05/horeca-top-100-2016-nummer-23-bilderberg-groep-101236366.

Salazar, D. A., Vélez, J. I., & Salazar, J. C. (2012). Comparison between SVM and logistic regression: Which one is better to discriminate? *Revista Colombiana de Estadística, 35*, 223-237.

Schmitt, P., Skiera, B., & Van den Bulte, C. (2011). Referral programs and customer value. *Journal of Marketing, 75*, 46-59.

Schmittlein, D. C., & Peterson, R. A. (1994). Customer base analysis: An industrial purchase process application. *Marketing Science, 13*, 41-67.

Shoemaker, S., & Lewis, R. C. (1999). Customer loyalty: the future of hospitality marketing. *International Journal of Hospitality Management, 18,* 345-370.

Tamaddoni Jahromi, A., Sepehri, M. M., Teimourpour, B., & Choobdar, S. (2010). Modeling customer churn in a non-contractual setting: the case of telecommunications service providers. *Journal of Strategic Marketing, 18,* 587-598.

Tamaddoni, A., Stakhovych, S., & Ewing, M. (2016). Comparing churn prediction techniques and assessing their performance: A contingent perspective. *Journal of Service Research, 19,* 123-141.

Tepeci, M. (1999). Increasing brand loyalty in the hospitality industry. *International Journal of Contemporary Hospitality Management, 11*, 223-230.

Van Leemput, V. (2016). *Predicting Churn in the Business-to-Business Cellphone Market* (Master Thesis). Retrieved from: Scriptiebank Tilburg University.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research, 218,* 211-229.

Vercellis, C. (2011). *Business intelligence: data mining and optimization for decision making*. John Wiley & Sons.

Weatherford, L.R., and S.E. Kimes. (2003). “A Comparison of Forecasting Methods for Hotel Revenue Management.” *International Journal of Forecasting, 19,* 401-415.

Wei, J. T., Lin, S. Y., & Wu, H. H. (2010). A review of the application of RFM model. *African Journal of Business Management, 4,* 4199.

Wu, S. I., & Lu, C. L. (2012). The relationship between CRM, RM, and business performance: A study of the hotel industry in Taiwan. *International Journal of Hospitality Management, 31*, 276-285.

Appendix

Appendix 1

Action Orders Matrix

Attribute	**Short Description**
ses_session_key	Unique key per data entry.
ses_session_type	Type of booking consisting of the levels hotel, conference, and dummy.
ors_person_key_guest_md5	Unique key per guest.
company_key	Unique key per business related booking.
ors_person_key_travelagent_md5	Unique key per travel agent making the booking (if relevant).
ses_property_key_md5	Unique key per Bilderberg establishment.
ses_session_group_key_md5	Unique key per group booking.
ses_datetime	Date and time of when a booking was placed.
ors_event_key	Unique identifier per booking.
ors_datetime_cancelled	Date and time of when a booking was cancelled.
ors_datetime_arrival	Date and time of check-in.
ors_departure arrival	Date and time of check-out.
ors_is_cancelled	Yes/no variable indicating whether a booking was cancelled.
ors_nr_of_rooms	Number of rooms per data entry.
ors_nr_of_adults	Number of adults per data entry.
ors_nr_of_children	Number of children per data entry.
ors_group_name	Name of group in Bilderberg system
ors_is_group	Whether the data entry is part of a group (yes/no).
ors_lead_time	Time between expected booking and last booking.
ors_length_of_stay	Length of stay per data entry.
ors_reservation_status	Arriving status of the guest(s).
ors_guarantee_status	Guarantee status of the guest(s).
ors_is_business_context	Yes/no variable indicating whether a booking was B2B.
ors_booker	Variable with three levels indicating whether the booker was a guest, a company or a travel agent.
ors_revenue	The sum of all revenue per data entry.
ors_revenue_1	All revenue related to accommodation income per data entry.
ors_revenue_2	All revenue related to food income per data entry.

ors_revenue_3	All revenue related to beverage income per data entry.
ors_revenue_4	All revenue related to other food and beverage income per data entry.
ors_revenue_5	All revenue related to telecommunications incomes per data entry.
ors_revenue_6	All revenue related to other income per data entry.
ors_revenue_7	All revenue related to leisure incomes per data entry.
ors_revenue_8	All revenue related to V.P.O. and non-trading expenses and income.
ors_revenue_9	All revenue related to ledgers and payments expenses and income.
ors_has_revenue	Yes/no variable indicating whether a data entry had revenue.

Note. All action orders attributes as they were received from Bilderberg.

Appendix 2

Company Specific Matrix

Attribute	**Short Description**
company_key	Unique key per data entry.
prd_type	Indicates whether the type of booking was a company (same input for all rows).
prd_first_name	Blank column indicating the first name of a B2C booker.
prd_last_name	Blank column indicating the last name of a B2C booker.
prd_last_ name_prefix	Blank column indicating the prefix of the last name of a B2C booker.
prd_full_name	String indicating the name of the company as it was entered into the database per data entry.
prd_birth_date	Blank column indicating the birthdate of a B2C booker.
prd_birth_date_year	Blank column indicating the birthdate of a B2C booker.
prd_email	String indicating the email address related to a booking (one can cover multiple data entries).
prd_phone1	String indicating the phone number related to a booking (one can cover multiple data entries).
prd_phone2	String indicating a second phone number related to a booking (one can cover multiple data entries).
prd_gender	Empty column indicating the gender of a B2C booker.
prs_membership_number	Empty column indicating the membership number of a B2C booker.
prs_membership_level	Empty column indicating the membership level of a B2C booker.
prs_last_transaction_date	Date and time of the last transaction per prs_company_corporate_key.
prs_company_corporate_key	Identifier of bookings made by the same company (not accurate).

Note. All company specific attributes as they were received from Bilderberg.

Appendix 3

Text manipulations of the ‘prd_full_name’ attribute.

```
## Setting the column to character
  data$prd_full_name <- as.character(data$prd_full_name)
## Setting al characters in the column to lower case
  data$prd_full_name <- tolower(data$prd_full_name)
## Removing general things.
  data$prd_full_name <- gsub("&", " and ", data$prd_full_name)
  data$prd_full_name <- gsub(" and ", " en ", data$prd_full_name)
  data$prd_full_name <- gsub("\\+", " en ", data$prd_full_name)
  data$prd_full_name <- gsub("\\.", "", data$prd_full_name)
  data$prd_full_name <- gsub("|\\'|\\(|\\)|\\-|\\,|\\/", "", data$prd_full_name)
  data$prd_full_name <- gsub("é", "e", data$prd_full_name)
  data$prd_full_name <- gsub("ä", "a", data$prd_full_name)
  data$prd_full_name <- gsub("à", "a", data$prd_full_name)
  data$prd_full_name <- gsub("ç", "c", data$prd_full_name)
  data$prd_full_name <- gsub("ü", "u", data$prd_full_name)
  data$prd_full_name <- gsub("ö", "o", data$prd_full_name)
  data$prd_full_name <- gsub("ë", "e", data$prd_full_name)
  data$prd_full_name <- gsub("aktiengesellschaft", "ag", data$prd_full_name)
  data$prd_full_name <- gsub("netherlands", "nederland", data$prd_full_name)
  data$prd_full_name <- gsub("holland", "nederland", data$prd_full_name)
  data$prd_full_name <- gsub("belgium", "belgie", data$prd_full_name)
  data$prd_full_name <- gsub("deutschland", "duitsland", data$prd_full_name)
  data$prd_full_name <- gsub("germany", "duitsland", data$prd_full_name)
  data$prd_full_name <- gsub("global", "international", data$prd_full_name)
  data$prd_full_name <- gsub("groep", "group", data$prd_full_name)
  data$prd_full_name <- gsub("university", "universiteit", data$prd_full_name)
  data$prd_full_name <- gsub("fryslan", "friesland", data$prd_full_name)
  data$prd_full_name <- gsub("bv|nv|vof", "", data$prd_full_name)
  data$prd_full_name <- gsub("  ", " ", data$prd_full_name)
## Removing all spaces
  data$prd_full_name <- gsub(" ", "", data$prd_full_name)
```

Note. All text manipulations were done in R.

Appendix 4

Features used in general model.

Feature	Short Description
prd_full_name	Unique identifier per company.
recency	Time between last transaction and end of calibration period per company.
observation_period	Time between start of calibration period and last transaction per company.
interpurchase	Average time between transactions per company in calibration period.
frequency	Frequency of transaction moments per company in calibration period.
ors_revenue	Aggregated revenue per company in calibration period.
churn	Binary variable indicating whether the company churned (1) or not (0).

Note. All features were aggregated and/or created with R.

Appendix 5

Features used in Bilderberg model.

Feature	Short Description
recency	Time between last transaction and end of calibration period per company.
observation_period	Time between start of calibration period and last transaction per company.
interpurchase	Average time between transactions per company in calibration period.
hotel	Frequency of transaction moments related to hotel bookings per company in calibration period.
frequency	Frequency of transaction moments per company in calibration period.
ors_revenue	Aggregated revenue per company in calibration period.
ors_nr_of_rooms	Aggregated numbers of rooms per company in calibration period.
ors_nr_of_adults	Aggregated number of adults per company in calibration period.
ors_length_of_stay	Aggregated number of days of stay per company in calibration period.
booker	Dummy variable indicating whether the company had more bookings done by the company itself (0), the guest himself (1), or a travel agency (2).
nr_locations	Amount of different Bilderberg locations visited per company in calibration period.
churn	Binary variable indicating whether the company churned (1) or not (0).

Note. All features were aggregated and/or created with R.