A Competing Risks Model Based on Latent Dirichlet Allocation for Predicting Churn Reasons

Dorenda Slof[a], Flavius Frasincar[a,*], Vladyslav Matsiiako[a]

[a]*Erasmus University Rotterdam, PO Box 1738, 3000 DR, Rotterdam, the Netherlands*

Abstract

Due to low switching costs and stiff competition, customer relationship management has become a central component in the marketing strategy of telecommunication service providers. Since the costs of acquiring a new customer are five times higher than the costs of maintaining an existing customer, telecommunication service providers are eager to reduce the churn rate. A solid understanding of customer churn behavior can help to address this problem. Reducing the churn rate can translate into significant revenue gains and might provide the edge to outperform the competitor. In this paper, we predict the propensity to churn for customers of a Dutch telecommunication service provider by employing a duration model. While predicting churn, we simultaneously predict the reason for which the customer churns, using a competing risks model. Since the telecommunication service provider has valuable textual data based on transcripts of calls between customers and the customer service center, we incorporate topics extracted from this textual data as variables in the competing risks model, by employing Latent Dirichlet Allocation (LDA). We find that we are among the first to incorporate LDA-based variables in a competing risks model. We compare four models and find that the models that have incorporated topic variables usually yield the best churn forecasts. Also, the investigated models beat the considered benchmark model, which is the model currently deployed at the telecommunication service provider.

Keywords: Churn, competing risk models, textual data, Latent Dirichlet Allocation

1. Introduction

Over the last two decades, the number of Internet users worldwide has increased at an impressive rate. According to “Miniwatts Marketing Group", around 4.5 billion people had access to the Internet on June 30, 2019, which tripled in the last ten years [1]. Due to the increase of Internet users worldwide, telecommunication service providers (TSPs) in developing countries benefit, as more and more people buy access to the World

*Corresponding author; tel: +31 (0)10 408 1340; fax: +31 (0)10 408 9162
Email addresses: dorendaslof@gmail.com (Dorenda Slof), frasincar@ese.eur.nl (Flavius Frasincar), matsiiako@ese.eur.nl (Vladyslav Matsiiako)

Wide Web. However, TSPs in developed countries are facing a completely different market, as around 75% of their citizens already have access to the Internet [1]. In order to maintain their price level, the TSPs often offer a variety of products and services to increase the customer base. On the other hand, the consumers of these countries are usually facing a limited number of highly competitive TSPs offering identical products and services, which makes them frugal consumers that tend to switch quickly between various TSPs. Indeed, according to [2], around 2% of the telecommunication company's customers churn every month.

Customer attrition is one of the major issues for a lot of industries but, in particular, to TSPs. In [2], the author estimates that 75% of the new subscribers have previously used the services of another TSP. Such a switching customer that leaves the current TSP is referred to as a churner. These churners are expensive for a TSP, as the costs of acquiring a new customer are five times higher than the costs of maintaining an existing customer [3]. Keeping in mind that TSPs spend a large amount of money on advertisements to gain new customers, TSPs aim to retain each customer in their customer base. In particular, many of the TSPs are not trying to sell as many products as possible to their existing customers or to lure the maximum number of new customers. Instead, they chose to adopt a retention approach with the goal to keep the potential churners for as long as possible and reduce the churn rate [4].

To improve retention, TSPs examine their customers' journey. We define the customer journey as the sequence of events that occurred during the contract period of a customer. Events such as calling the customer service center, visiting the website of the TSP, or ordering an extra product could all be part of the customer journey. The content of the calls with the TSP, further referred to as the customer service data, is expected to be of particular importance in the customer journey [5]. We hypothesize that the customer journey of a particular customer will provide signals about the customer's propensity to churn and the reason for that.

There exists a considerable amount of research directed toward predicting the probabilities of customers to churn [6]. At the same time, there is almost no research that investigates the prediction of customer churn reasons. Furthermore, none of the previous works investigate these two issues simultaneously, even though this could unlock new practical applications. For example, if we are able to identify potential churners in the earliest stage possible, the associated TSP can contact these customers and try to prevent them from leaving before it is too late. If, in addition to that, we predict the reason for which a customer churns, the TSP could come up with the optimal personalised strategy for preventing this customer from leaving. Therefore, the research question of our study is posed as follows: "How can we predict the customers' likelihood to churn and the reasons for that incorporating customer service data?"

For preventing future churn the TSP needs to know which variables in the customer journey influenced already churned customers, such that they can control for these variables in the future. The topics of customer service calls, price of the contract, webpage visits, and the number of failures are all examples of such possible

variables. Gaining insight into the customer journey has the potential to reduce the customer churn rate and improve company-customer relations, both leading to an increase in revenue.

Even though some researchers, like [5], have already looked into incorporating the text information (e.g., customer e-mails sent to the service center) into churn models, to our knowledge, no research has been done on using the topics present in text data (in our case, support requests) in duration models. In addition, we have also researched the incorporation of textual variables into a competing risks model, by applying Latent Dirichlet Allocation (LDA). Even though it has not been previously investigated, identifying the topics of user support requests might be the most important factor in predicting churn as it points out the reason for which the customer is unsatisfied. We have shown that this approach helps to boost the performance of the considered models. Using our results, managers of telecommunication companies will be able to improve their churn models which should help them increase customer retention and fight churn better.

This paper is structured as follows. First, we discuss the related work. Then, we give an overview of the empirical data used for this research. The applied methods are explained next. We end this paper with a discussion of the results of the applied methods and propose topics for future work.

2. Related work

Customer retention continues to be a hot topic among marketeers and managers in the telecommunication market [7]. Recent studies focus on churn probabilities and the prediction of the customer lifetime value [8], the allocation of resources to customer retention and acquisition [9], effects of information and communication technology (ICT) service and its complementary strategies on customer loyalty [10], and the formation of financial reports and communication with management regarding churn [11]. Overall, we find that the majority of the research on churn is focused on *which* customers are going to leave. In addition, researchers also investigate the moment *when* a customer ends the contract and on the duration of the contract.

Another stream of research is focused on *why* a customer terminates his/her contract. For example, [12] states that determining the most important causes of customer churning can improve the understanding of predictive models while saving a lot of expenses and computational time. As is found by [13], product line breadth and quality reduce customer switching and may also reduce customer attrition. Specifically, for instance, using Markov logic networks, [14] points out that word-of-mouth can significantly influence customer churning decisions. As discussed by [7], one can classify churners into three classes, also referred to as risks: Value, Personal, and Non-Pay. The Value risk includes anything related to the sentiment with respect to the company, such as response to poor service, price of the product or service, and the quality of the product or service. These risks are all under control of the company and are therefore defined as Controllable churn, as

described by [15]. The Personal risk includes all risks that are outside the control of the company, such as migration or death. The Non-Pay risk includes all risks related to non-payment, such as abuse or theft of the product or service. On all occasions of Non-Pay churn, the company makes the decision to end the contract of the customer. The Personal risk and the Non-Pay risk combined can be regarded as Uncontrollable risk.

Predicting the reason for which a contract is terminated has extensive managerial implications. As mentioned by [8] and [11], the firm's predicted aggregate churn rate is sufficient to evaluate the financial health of an organization and to estimate the value of the customer base. However, a manager who has the responsibility to reduce the churn rate can respond to this type of knowledge, by giving customers with a high propensity to churn the right offer [16]. For instance, if we expect that a customer is likely to churn due to relatively high prices, the manager could offer this customer a more appealing product for a better price to retain this customer. If the manager is able to do so before the customer was planning to terminate his/her contract, the manager will reduce the churn rate.

In the field of customer churn prediction, authors typically adopt a data mining approach to predict which customers are likely to churn [6, 17–20]. Based on historical data, a model is trained to classify customers as future churners or as future non-churners. Modern classification techniques such as random forests, support vector machines, and neural networks are often adopted to predict churn [17, 21–23]. Although many studies compare the performance of various classification techniques, there is no consensus yet about which classification technique performs best for customer churn prediction [24, 25].

Among others, [26] compares the performance of various classification and econometric models for churn prediction. Despite the good predictive power of classification models, a drawback of adopting them to predict customer churn is that most of them only result in classification of churners and non-churners. Yet, it is desired to have the results in terms of timestamps when customers churn to support customer triage. The second drawback is that for most classification techniques the effects of the variables on the churn likelihood are hard to interpret [27, 28]. Although many studies focus on extraction rules to increase the comprehensibility of the effects of variables on the propensity to churn, the improvement is only marginal [29].

In the more specific field of prediction of the moment *when* a customer churns, authors often adopt duration models for predicting the propensity to churn [30, 31]. A duration model provides the length of time spent in a given state before the transition to another state [32]. For the TSP, a transition occurs to an exit state when a customer terminates the contract. The key concept in duration modeling is the hazard rate or hazard function, which is the probability that a contract is terminated after a certain number of months, given that it is not terminated yet [32]. Since the selection of the form of the hazard function is rather arbitrary, [33] proposes the Cox model with the unspecified hazard function. In this case, it is not need to select an underlying distribution of the hazard function, while the purpose of the model remains valid.

A duration model seems to fit well with the objectives of our work due to several reasons. First, the probability to churn is one of the results of the duration model. Also, the interpretation of the parameter estimates of the duration model is comprehensible and straightforward [32]. Last, the duration model specifically predicts *when* a customer is likely to churn.

A duration model only allows for a transition from one state to another state. To predict the reason for which a customer churns, we aim for a predictive model that considers more than one exit state. This functionality can be found in a competing risks model, which can be regarded as an extension of the duration model. A competing risks model takes multiple exit states (or risks) into account, where different risks race to decide which risk will trigger the eventual churning act [7]. [34] claims that most studies on competing risks model development are concentrated in natural sciences. A recent overview of competing risks models in biostatistics is given by [35] and in medical research by [36]. Competing risks models are also suitable in many economic applications. To analyze unemployment, [37] and [38] propose several competing risks models. [39] and [40] apply a competing risks model to predict the inter-purchase time for a household, which turns out to describe the inter-purchase time adequately. [41] proposes a competing risks model to explore the duration of participation in the market for high tech firms.

In addition, there is some research which proves that adding unstructured textual data for the issue of churn prediction may be useful [42]. As [7] mentions, incorporating explanatory variables such as calls with the customer service center, online activities, or any touchpoints the TSP has had with the customer, will have considerable practical value. Among others, [31] incorporates touchpoints into a duration model, and shows that links can be found between the propensity to churn and variables capturing customer service experience. Furthermore, [5] and [43] prove that variables capturing the content of a conversation with the customer service center boost the predictive performance of their proposed models. Additionally, [44] investigates incorporating textual data into the customer churn forecast models and show that it significantly improves the performance of the prediction models.

To capture the content of the conversations with the customer service center in a variable, we suggest detecting the underlying topics of each conversation. For this purpose, we propose a topic model, which is a probabilistic model for uncovering the underlying semantic structure of a document collection, such that each document can be assigned to latent topics [45]. Topic models can extract surprisingly interpretable and useful structures without any explicit understanding of the language by the computer [46]. The most used model types for topic modelling are (probabilistic) Latent Semantic Analysis [47] and Latent Dirichlet Allocation (LDA) [48]. The latter one, which is also the most recent one, will be used for the purposes of our research.

The goal of LDA is to find topics for each document in a document-collection automatically. We have chosen LDA as it is one of the best topic modeling techniques showing superior results to Latent Semantic

Analysis (LSA) and probabilistic LSA (pLSA) according to [48]. Incorporating the LDA generated topics of each conversation into a predictive model has been extensively studied. [49] applies LDA on Twitter data for predicting future hit-and-run crimes, which outperforms a baseline model that predicts hit-and-run crimes uniformly across all days. The completion time of crowdsourced tasks has been studied by [50], who use LDA for detecting which groups of tasks are picked up faster than other groups of tasks. In political research, [51] and [52] use LDA to examine the relationship between legislative text and legislative sentiment. [53] employs LDA to examine the relation between written and spoken words, and political conflicts. In social sciences, both [54] and [55] apply LDA to examine whether a member in an online support group was emotionally supported. They demonstrate that emotionally supported members are less likely to drop out of the online group. LDA is employed in economic applications as well. Among others, [56] applies LDA to news content, improving their predictions of stock market behavior.

The novelty of our research can be explained as follows. As previously mentioned, [5] incorporates the content of e-mails of customers sent to the customer service center into a churn prediction model. However, to our knowledge, we are among the first to incorporate the content of conversations between customers and the customer service center into both a duration model and a competing risks model. Moreover, we pioneer the incorporation of textual variables in general into a duration model and a competing risks model, by applying LDA. In our case, the textual variables represent the customer service data which we hypothesise to be of significant value when predicting churn. Additionally, incorporating touchpoints in a competing risks model is a novelty by itself when predicting the reason for which a customer churns.

3. Empirical data

In this study, we have information on 589,985 current customers and 201,815 churners of a Dutch TSP that has been gathered from January 2014 until March 2016. We have two sources of information: offline data, and online data. Although most of the data is offline information (e.g., demographical data, financial data, records of the customer service calls, etc.), some customers also reveal parts of their online behavior (e.g., visits to webpages of the TSP). Offline data is available for all customers, while the online data is only available for customers that logged in to their personal environment at the website of the TSP after April 2015. Notice that the timespan of the online dataset is only a subset of the timespan of the offline dataset.

In addition, it should be mentioned that the data obtained is deemed reliable. In fact, there is no missing or outlier values. All the individuals which had certain data points missing or falsified (the age is higher than physically possible) were omitted from the dataset. Next to that, we believe that the validity of the data is also on a high level which means that both demographic and behavioural characteristics of users represent

the reality (whenever a user triggers a certain event, such as filing a complaint or visiting the website, it becomes a fact that is immediately known to the TSP).

3.1. Offline Data

The offline dataset consists of the variables shown in Table 1. This table also shows the descriptions of the variables and their types (e.g., textual variable, categorical variable, etc.). For each category of the categorical variables, we create a dummy variable that takes on the value 1, if a certain category applies to the customer, and the value 0, otherwise. For each categorical variable, we discard one of the dummy variables, to avoid multicollinearity into the proposed models. Multicollinearity occurs when two or more dummy variables are highly correlated, that is when one variable can be predicted from the others.

As can be seen in Table 1, the offline dataset can be divided into six main categories of variables. For the first category, we consider the *customer characteristics*, which consist of variables such as the unique identifier, zip code, and city. The second category contains the *product characteristics*, which consist of the products and services purchased by the customer, the price of the contract, and the channel where the products and services are purchased. *Billing information* is added into the third category, containing information about the non-payments of the customer. Fourth, we consider *customer service information*, which consists of the contacts the customers have had with the customer service center. The customer service information mainly contains textual data. In fact, [57] finds that including textual data in the model for customer churn prediction does significantly improve its performance. The *failure information* is the fifth category, which contains the date and the location of the failures. Finally, we consider *churn characteristics*, consisting of variables such as the date of churn, and the reason for which a customer churns. The variables of the *customer characteristics* and the *product characteristics* are all constant over time, while the variables of the *billing information*, the *customer service information*, and the *failure information* differ over time.

Table 1 Offline data

Variable	Description	Type[1]
Customer characteristics		
Customer ID	The identity of the customer.	T
Zip code	The zip code of the customer.	T
Fiber	An indicator value for whether the customer has a fiber connection, or not.	C
City	An indicator value for whether the customer lives in a city, or not.	C
Mosaic group[2]	The mosaic group to which the customer belongs.	C
Product characteristics		
Play name[2]	The informal name of the products and services the customer owns.	C

Table 1 – continued from previous page

Variable	Description	Type[1]
Channel[2]	The channel where the first product or service is purchased.	C
Price	The price of all products and services the customer owns.	C
Value Added Services[2]	The extra services the customer has on top of the regular contract.	C
Billing information		
Warning page (WP)	An indicator value for whether the customer has seen a warning page online, which will be shown when the customer's payments are delayed.	C
Warning page, soft disconnect (WPSD)	An indicator value for whether the customer has experienced a soft disconnect when he/she did not respond to the warning page: the products and services are temporarily unavailable.	C
Customer service data		
Customer ID	The identity of the customer.	T
Contact date	The date the TSP either called the customer, or the TSP was called by the customer.	D
Subject	The subject of the contact with the customer.	T
Comments	A small summary of the contact with the customer.	T
Response	The response of the customer service employee to the customer.	T
Failure information		
Failure date	The date of the failure.	D
Zip code	The zip code where the failure occurred.	T
Churn characteristics		
Date contract started	The start date of the contract.	D
Date termination received	The date the customer terminated its contract.	D
Reason termination	The reason for which the customer terminated its contract.	C

[1]We distinguish the data types: Textual (T), Categorical (C), and Date (D). [2]The list of existing categories can be found in Appendix A.

3.1.1. Customer Service Data

The customer service data contains information about the 5,542,401 calls between customers and the customer service center of the TSP, from October 2014 until March 2016. Note that the timespan of the customer service data is only a subset of the timespan of the data about customers and churners. For each reached customer, the customer service employee records the subject of the call with a short summary. Table 2 shows some examples of summaries of calls between the TSP's customers and its customer service center.

Table 2: Examples of summaries of calls

Customer ID	Summary
1	Customer says he is not going to pay the bill of December.
2	Interactive Television is still not connected, mechanic is sent.
3	Explained how the laptop should be connected to the Wi-Fi.
4	Question: we do not use Interactive Television anymore. Can we terminate that part of our contract?

Primarily, we examine the effect of the number of contacts per month the customer has had with the customer service center on the propensity to churn. We incorporate this variable into the proposed models. However, we also expect that including the content of the calls in our proposed models increases the performance of the models, both for predicting the propensity to churn and for predicting the reason and the propensity to churn jointly. For example, when regarding the topic of the conversation of customer 1 from Table 2, this customer is likely to churn due to the Uncontrollable risk, since this conversation is about non-payment. If we would only incorporate the number of calls between the customer and the customer service center in the proposed models, this behavior cannot be captured. Moreover, we see that the customer service center was able to help customer 3, who is less likely to churn after this conversation since they feel satisfied with the help of the customer service center employee. Again, if we only incorporated the number of calls between the customer and the customer service center, this nuance could not have been captured. In order to include the subject and the summary of a call, we should convert the textual variables to the categorical ones. For this purpose, we implement an LDA model, which will be discussed in Section 4.3.

Note that we never incorporate variables about the number of calls and about the topic of the calls in the proposed models at the same time, since these types of variables are multicollinear.

3.1.2. Churn Characteristics

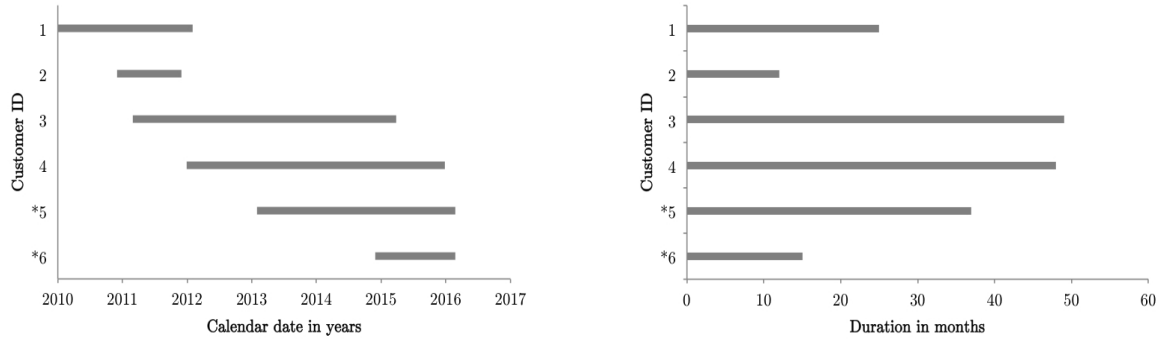
The churn behavior of all customers is available in the churn characteristics. As shown in Table 1, we exactly know when and why a customer terminated its contract. Although the exact date of churn is given for each customer that has churned, we only use the month in which the customer churns. The main reason is that only 0.04% of all customers churn each day. Training a model on such an unbalanced dataset will give unreliable results. By aggregating the period in which a customer could churn to months, approximately 1.3% of all customers churn each month, which has the potential to generate more reliable results.

The dependent variable for our proposed models consists of two parts: the *duration* variable, and the *status* variable. The *duration* variable represents the difference in months between the start date and the termination date of the contract. The relation between the duration of the contract and the calendar dates of the contract can be seen in Figure 1. Figure 1(a) shows the calendar months in years of being in contract for six random customers, while Figure 1(b) shows the duration of each contract, where customers with an asterisk did not churn yet.

The *status* variable depends on the purpose of the model, which could either be to predict churn or to predict both churn and the reason for which a customer churns. For the former purpose, that is, for the duration model, the value of the *status* variable is equal to zero for current customers and equal to one for customers who already terminated their contract. Similarly, for the latter purpose, that is, for the competing

risks model, the value of the *status* variable is zero until the customer churns. Then, the value of the *status* variable depends on the reason for which the customer churns, where each reason is expressed as a different integer.

In the churn characteristics, we identify 59 distinct reasons why the product or service is terminated. These reasons are all recorded by the customer service center's employees. Since the performance of a competing risks model declines when the number of competing risks increases, we combine these reasons into three classes. Some of the reasons are self-reported (e.g., low Internet speed, disappointing service, etc.), while others are determined more objectively (e.g., migration, non-payment, etc.). Most risks can be distributed among the classes Value, Personal, and Non-Pay, which were explained in Section 2. Since the customers potentially ought to the Value class can be controlled in some way, we rename this class to the Controllable class. Furthermore, since the TSP is unable to control customers potentially ought to the Personal class or the Non-Pay class, we aggregate these classes into the Uncontrollable class. Some risks are reported as unknown, which makes it unable to classify these risks into one of the aforementioned classes. We assume that customers in the Unknown class are careless customers since they did not provide a reason for which they terminated their contract with the TSP. Hence, we form a third Unknown class containing customers who did not provide a reason for which they churned. Table 3 shows the different competing risks with their descriptions.



(a) Calendar months in contract

(b) Duration of the contract

Figure 1: Duration variable

Table 3: Description of competing risks

Risk	Description
Controllable	All risks which are in control of the TSP, such as churn due to high prices, bad services, or a better offer of a competitor.
Uncontrollable	All risks which are out of control of the TSP, such as migration or death.
Unknown	The risk that contains all customers who did not provide a reason for which they churned.

3.2. Online Data

The online dataset consists of the variables as shown in Table 4, either gathered from the website of the TSP or the TSP's family companies. These family companies (fTSP) offer similar products and services as the TSP, but could differ on their service levels, their offered products and services, and the price of their contracts. Note that the online data is not constant over time.

Table 4: Online data

Variable	Description	Type[1]
Cookie ID	The online identity of a customer.	T
Time	The date and time the user with the cookie ID visited a website of either the TSP or an fTSP.	D
Other data	The *url* of the visited webpage.	T

[1]We distinguish the data types: Textual (T), and Date (D).

We have rather detailed information about the online activities of some customers. Online activities are clicks on the website of the TSP or the fTSP. Deduced from these online activities, we add variables to the variable set which count the number of activities on the website of the TSP or the fTSP for each month, starting in May 2015. We expect that online customer behavior will provide information about churn and the reason for which a customer churns since customers potentially visit the website of the TSP when they have churn-related questions. Moreover, we expect that customers visiting the website of a sister company are searching for another offer, which implies churn at the TSP.

4. Methodology

As discussed in Section 2, we model the propensity to churn using a duration model. For jointly modeling the propensity to churn and the reason for which a customer churns, we adopt a competing risks model. As mentioned in Section 2, a competing risks model can be regarded as an extension of a duration model. Moreover, we employ LDA on the content of the conversations between the TSP's customers and its customer service center, to capture the topics of this textual data into a variable.

4.1. Duration Model

A duration model is used to model the length of time spent in a given state before transitioning to another state, as described by [32]. Denote by T a continuous random variable, also referred to as the duration variable, where T represents the time in months a customer is in a contract. Then, we define the hazard function $\lambda(t)$ as the instantaneous probability to churn at time t, given that the contract is not terminated yet on time t. This hazard function is equal to:

$$\lambda(t) = \lim_{\delta \to 0} \frac{P\left(t \leq T < t + \delta \mid T \geq t\right)}{\delta} \tag{1}$$

Since we also aim to estimate the effects of the explanatory variables on the instantaneous probability to churn, we parameterize the hazard function as follows:

$$\lambda\left(t \mid X(t-1)\right) = \lambda_0(t)\exp\left(X(t-1)\beta\right) \tag{2}$$

where $\lambda_0(t)$ is the baseline hazard function on time t, which is identical for all individuals but differs over time t. Furthermore, $X(t-1)$ is the $N \times k$ matrix with explanatory variables on time $t-1$, with N equal to the number of customers the TSP has ever had, k the number of explanatory variables, and β the $k \times 1$ vector with model parameters. The baseline hazard function $\lambda_0(t)$ describes the instantaneous probability to churn for individuals with $X(t-1) = 0 \ \forall \ t$. Both the baseline hazard function $\lambda_0(t)$ and the parameters β are unknown, and should respectively be selected beforehand or estimated.

We estimate the parameters β by applying the extended Cox model, as described in [32]. Then, we are also able to predict the instantaneous probability to churn $\hat{\lambda}\left(t \mid X(t-1)\right)$ for each customer at time t.

4.2. Competing Risks Model

We generalize the concepts of the duration model from a model with one exit state to a model with three exit states, to jointly predict churn and the reason for which a customer churns. The exit states are the reasons for which a customer could churn, given by three risks: the Controllable risk, the Uncontrollable risk, and the Unknown risk. By applying a competing risks model on the data, we simultaneously model transitions to each of the three exit states.

We describe the set of risk types as $r = \{1, 2, 3\}$, where each exit state is one of the risks. The three risk types in the risk set r are respectively the Controllable risk, the Uncontrollable risk, and the Unknown risk. In total, we know four states: $\{0, 1, 2, 3\}$, where state 0 is the initial state, that is, the state of being a paying customer. Note that we assume independent risks since we assume that a customer cannot be at high risk for two risks simultaneously.

Denote with T_j a continuous random variable for risk $j \in r$. T_j represents the time in months a customer is in contract until churn, due to risk j. Then, the unparameterized hazard function is given by

$$\lambda_j(t) = \lim_{\Delta t \to 0} \frac{P\left(t \leq T_j \leq t + \Delta t \mid T_j \geq t\right)}{\Delta t} \tag{3}$$

where λ_j is the unparameterized hazard function of risk j, for $j = 1, 2, 3$. To estimate the effects of the explanatory variables on the length of the customer's contract, and on the reason for which a customer churns,

we parameterize the hazard function as

$$\lambda_j(t \mid X_j(t-1)) = \lambda_{0j}(t) \exp\left(X_j(t-1)\beta_j\right) \tag{4}$$

where $X_j(t-1)$ is the $N \times k$ matrix with explanatory variables for risk j on time $t-1$. Note that the set of explanatory variables could differ per risk. Moreover, the baseline hazard function λ_{0j} and the $k \times 1$ vector with model parameters β_j are also specific for risk type j. As before, $X_j(t-1)$ represents the explanatory variables for risk j measured on time $t-1$, to be able to predict the instantaneous probability to churn on time t.

We estimate the parameters β_j by applying the extended Cox model again, as described in [32]. Then, we are also able to predict the instantaneous probability to churn due to risk j, $\hat{\lambda}_j\left(t \mid X(t-1)\right)$, for each customer at time t.

4.3. Latent Dirichlet Allocation

As mentioned before, the TSP's customer service center employees record all conversations they have had with their customers. By applying LDA on the recorded conversations, we can determine the topic of each conversation. LDA models each call between the customer and the customer service center as a mixture over topics. We assume there exist K topics, where each topic is a Dirichlet distribution over a vocabulary with V terms, with parameter $\eta_k = \{\eta_{kv}\}$ for $k = 1, ..., K$ and $v = 1, ..., V$ with prior $\beta = 0.1$. For conversation m, we draw a mixing proportion $\theta_m = \{\theta_{mk}\}$ over $k = 1, ..., K$ topics from a Dirichlet distribution with prior $\alpha = 0.1$. For the nth word in conversation m, a topic z_{mn} is drawn from a Multinomial distribution with parameter θ_m, where topic k is chosen with probability θ_{mk}. Then, word w_{mn} is drawn from the term distribution η_{kv} given $k = z_{mn}$. Thus, term v for topic k is chosen with probability η_{kv}. Then, the joint distribution over all parameters is given by:

$$p(\mathbf{w}, \mathbf{z}, \theta, \eta \mid \alpha, \beta) = p(\eta \mid \beta)\ p(\theta \mid \alpha)\ p(\mathbf{z} \mid \theta)\ p(\mathbf{w} \mid \mathbf{z}, \eta) \tag{5}$$

where $\mathbf{w}$ is the matrix with assigned words, $\mathbf{z}$ is the matrix with assigned topics, $\theta = \{\theta_{mk}\}$ for $m = 1, ..., M$ and $k = 1, ..., K$, and $\eta = \{\eta_{kv}\}$ for $k = 1, ..., K$ and $v = 1, ..., V$. By estimating the posterior distribution $p(\mathbf{z}, \theta, \eta \mid \mathbf{w}, \alpha, \beta)$ we can determine the topic of each call. This posterior distribution is estimated by applying the collapsed Gibbs sampler, as explained in [58].

5. Results

We apply the duration model and the competing risks model described in Section 4 on the dataset described in Section 3. We use the statistical language R for this purpose. We used the `survival` package to model and predict the churn models. The `topicmodels` package is used to generate the LDA model. As discussed, we have data that has been gathered from January 2014 until March 2016. For the churn models, we split this dataset into two subsets, to obtain a training set and a test set. The training set consists of 26 months, starting from January 2014 until February 2016. The test set consists of one month, which is March 2016. We compare and evaluate churn predictions of the duration model and the competing risks model. For the LDA model, we have a training set consisting of the conversations between the TSP's customers and the customer service center, from January 2014 until February 2016. The topics for the test set are determined using a model built on the training set. We compare and evaluate the duration model and the competing risks models both without and with the LDA generated topics as variables.

Before we discuss the performances of the proposed churn models, we will evaluate the results of the LDA model. These results will help to determine the topics of the calls between a customer and the customer service center of the TSP. Then, the topics can be incorporated into the proposed models as variables. Second, we discuss the in-sample performance measures of our churn models. Thereafter, we show the out-of-sample performance results of the churn models. Last, we give and interpret the parameter estimates of the churn model.

5.1. LDA Model

At first, we should learn the optimal number of topics for our LDA model from the training data generated for LDA. In order to select the optimal number of topics K in our LDA model, we should compare multiple LDA models differing on K using a certain performance measure. Though, since LDA models are unsupervised, not many evaluation methods are applicable. [48] propose to evaluate an LDA model on the perplexity of a held-out test set. The perplexity is a measurement of how well a probability model predicts a sample. A convenient property of the perplexity is that it decreases when the likelihood of the model increases. Therefore, a lower perplexity indicates higher performance of the LDA model.

The perplexity of a held-out test set of an LDA model with K topics could be calculated once for the whole dataset. However, to validate the perplexity results and to generalize the results to an independent dataset, we execute the k-fold cross-validation technique. This technique randomly splits the original training set into k equal-sized subsets. Then, $k-1$ subsets are combined to form the training set, while the kth subset forms the test set. This process is repeated k times, where each subset is part of the training set $k-1$ times, and part of the test set once. Then, the perplexity is reported for the test set.

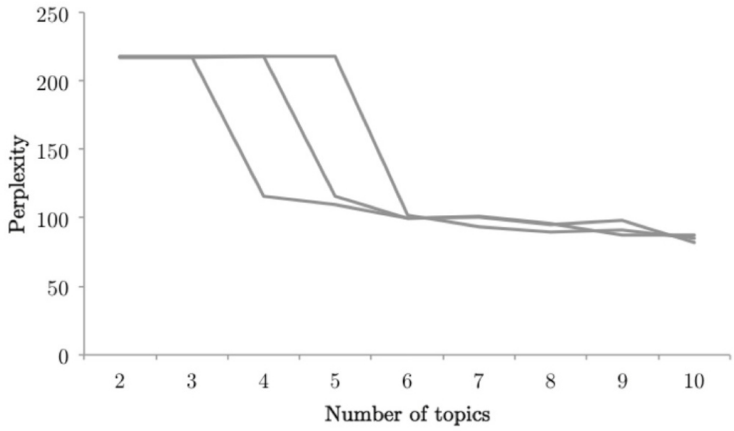


Figure 2: Perplexity of LDA models

Since our dataset of calls between customers and the customer service center consists of more than 5,000,000 observations, we choose $k = 3$ folds for executing the cross-validation technique. Figure 2 shows the results in terms of the perplexity of the 3-fold cross-validation technique, which is applied to LDA models differing on their number of topics K. As mentioned before, we select the optimal number of topics K by choosing a model with a small perplexity, compared to the other LDA models. In Figure 2 it can be seen that the perplexity of each fold does not decrease considerably after six topics. Therefore, we select $K = 6$ topics to be the optimal number of topics for this particular dataset.

It has to be noted that because of the size of our dataset, using the 3-fold cross-validation already requires a substantial time investment. Therefore, using 5-fold or 10-fold cross-validation would be even more time-consuming. This, in turn, would make it harder to apply in practice.

After selecting the optimal number of topics, we generate the final LDA model with the optimal number of topics K, applied on the training set generated for LDA. From the results of this model, we extract the topic of each conversation between a customer and the customer service center. One of the results can be seen in Table 5, which shows the six topics with the ten most likely terms for each topic. For each term, the probability that a term belongs to that topic is shown as well.

Table 5: LDA terms per topic with corresponding probabilities

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
abuse	**question**	**use**	**administr.**	**status**	**pay**
(0.096)	(0.135)	(0.122)	(0.059)	(0.140)	(0.094)
technique	**information**	**usage**	outbound	**order**	**explain**
(0.072)	(0.129)	(0.068)	(0.049)	(0.107)	(0.081)
outbound	save desk	inbound	call	Internet	**bill**
(0.062)	(0.090)	(0.065)	(0.046)	(0.067)	(0.080)
modify	sale	login	CC unit	deliver	non paid
(0.056)	(0.076)	(0.062)	(0.030)	(0.050)	(0.077)
save desk	customer	Internet	address	inbound	check
(0.054)	(0.071)	(0.057)	(0.030)	(0.042)	(0.074)

It can be seen in Table 5 that topic 1 is probably about abuse of the products and services of the TSP, and about technical problems. Note that the probability that the first term belongs to topic 1 is more than 9%. Topic 2 represents questions of existing customers, as the term *question* and the term *information* are likely to belong to topic 2. These terms have a substantially higher probability to belong to topic 2, than all other terms in topic 2 have. Topic 3 characterizes calls about the *usage* of the products and services. Table 5 also shows that topic 4 is related to *administrative* questions. Though, the probabilities that the terms of topic 4 actually belong to topic 4 are rather small in comparison with the probabilities of terms assigned to other topics. Topic 5 is about the status of customer's orders since the terms *status* and *order* have a high probability to belong to this topic. Finally, topic 6 is about explaining *payments* and *bills*.

After generating the final LDA model, we determine the posterior probability of belonging to a specific topic, for each call between a customer and the customer service center. We assume a call belongs to a certain topic when the posterior probability is higher than $\frac{1}{K}$, where $K = 6$ is the number of topics. We assign 1 to the topic variable of the call if the posterior probability of belonging to that topic is higher than $\frac{1}{6}$. Moreover, we assign 0 to the topic variable of the call if the posterior probability of belonging to that topic is lower than $\frac{1}{6}$. Note that we could assign more than one topic to each call between a customer and the customer service center of the TSP. In the end, we incorporate the topics of the calls as binary variables in the proposed models, for each customer and each time period.

5.2. Comparison of Churn Models

To compare the in-sample performances of the different churn models, we consider the log-likelihood (LL) value of the model, the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). These in-sample performance measures indicate the goodness-of-fit of the models on the training data. The performance measures are given by:

$$\mathrm{LL} = \ln(L) \tag{6}$$

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{7}$$

$$\mathrm{BIC} = \ln(N)k - 2\ln(L) \tag{8}$$

where L is the maximum value of the likelihood function from the churn model, k is the number of explanatory variables in the churn model, and N is the number of observations. While a higher LL value indicates higher performance of the model, a lower AIC or BIC indicates higher performance of the model. We select the model with the highest LL value, and with the lowest AIC and BIC.

Table 6 shows the in-sample performance of the proposed duration models in terms of the LL, AIC and

Table 6: In-sample performance of duration models

	Without topic variables			With topic variables		
	LL	AIC	BIC	LL	AIC	BIC
Churn risk	-2,520,456	5,041,054	5,041,782	**-2,469,176**	**4,938,493**	**4,939,221**

BIC values based on the full training dataset. Note that we refer to the churn risk when we discuss the duration model. The left-hand side performance measures in Table 6 correspond to the duration model without topic variables, while the right-hand side performance measures correspond to the duration model with topic variables. The highest LL value and the lowest AIC and BIC are shown in bold font. All values in Table 6 show that the duration model with incorporated topic variables performs better than the duration model without incorporated topic variables. Therefore, we conclude that incorporating topic variables into this duration model improves the in-sample performance.

Table 7: In-sample performance of competing risks models

	Without topic variables			With topic variables		
Risk	LL	AIC	BIC	LL	AIC	BIC
Contr.	-2,822,791	5,645,724	5,646,459	**-2,790,314**	**5,580,771**	**5,581,506**
Uncontr.	**-1,106,932**	**2,214,007**	**2,214,680**	-1,108,820	2,217,781	2,218,454
Unknown	-4,234,255	8,468,651	8,469,415	**-4,200,905**	**8,401,952**	**8,402,716**

The in-sample performance of the competing risks models based on the full training dataset can be found in Table 7. If we compare the left-hand side models with the right-hand side models, we observe that the models with topic variables for both the Controllable risk and the Unknown risk perform better than a model without topic variables. For the Uncontrollable risk, the model without topic variables is the best performing model. Probably, customers who churn due to the Controllable risk or due to the Unknown risk tend to call the customer service center with a specific problem, while customers who churn due to the Uncontrollable risk do not call the customer service center with a specific problem. Since customers who churn due to the Uncontrollable risk churn due to for example death, the topics of the calls with the customer service center do not improve the Uncontrollable risk model. For customers who churn due to the Uncontrollable risk, the number of calls between the customer and the customer service center is more relevant than the topics of the calls.

In the remainder of this paper, we will always use the best performing models. For example, when we speak of the Controllable risk model, we refer to the Controllable risk model with topic variables.

5.3. Predictive Performances

We solely evaluate the out-of-sample predictive performance of our best performing models. We compare the predictive performances of our proposed models with the predictive performances of a benchmark model

(BM). This benchmark model randomly selects $\bar{N}$ customers as churners for March 2016, where $\bar{N}$ is the average number of churners per month between January 2014 and February 2016. For each competing risk the benchmark model randomly selects $\bar{N}_C$, $\bar{N}_{UC}$ or $\bar{N}_{UK}$ customers as churners for March 2016, respectively the average number of churners per month due to the Controllable risk, due to the Uncontrollable risk or due to the Unknown risk between January 2014 and February 2016. This benchmark model is the currently employed predictive model at the TSP.

To determine the predictive performance for each model, we first consider the precision at r. To calculate the precision at r, we order the predicted instantaneous probabilities to churn of all customers from high probability to churn, to low probability to churn. This results in the ordered churn list. Then, the precision at r can be written as:

$$P@r = \frac{TP}{r} \tag{9}$$

where TP are the true positives, which is the number of correctly predicted churners in the first r observations from the ordered churn list, and r is the number of considered observations. We choose $r \in \{100, 500, 1000, \bar{N}\}$, since the TSP maximally acts on the first $\bar{N}$ churners from the ordered churn list.

Table 8: Out-of-sample performance of duration model and competing risks model

	BM	P@100	P@500	P@1000	P@$\bar{N}$	$\bar{N}$
Churn risk	0.014	0.200	0.146	0.120	0.073	7745
Controllable risk	0.004	0.050	0.046	0.038	0.034	2783
Uncontrollable risk	0.002	0.050	0.042	0.040	0.033	1245
Unknown risk	0.011	0.050	0.054	0.051	0.045	3717

Table 8 shows the precision at r for the benchmark model, just as the precision at r for each proposed model, and the average number of churners for each risk. All models outperform the benchmark model in terms of the $P@r$ value, which implies that the TSP should use the predictions of our proposed models for targeting customers who are likely to churn, instead of continuing their own targeting approach. For the Churn risk, we see that 20% of the first 100 customers on the ordered churn list are correctly predicted. Table 8 also shows that the $P@r$ value for the Churn risk decreases when r increases. The $P@r$ values of the competing risks are rather small. For example, for customers who churn due to the Controllable risk, we correctly predict just 3.4% of the 2783 high-ranked potential churners from the ordered churn list. This implies that all other customers are incorrectly predicted as churners, while they were actually no churners. A potential reason is that only 0.4% of the customers of the TSP churn due to the Controllable risk each month. Correctly predicting such a small number of churners has always been a difficulty in predictive churn models. For the Uncontrollable risk and the Unknown risk, we observe similar patterns. The $P@r$ values of

both risks are higher than the $P@r$ value for the benchmark model, but they are still rather small. Again, a potential reason is that only a small percentage of customers of the TSP churn due to the Uncontrollable, and the Unknown risk each month, which is a difficulty in predictive churn models.

We also determine the predictive performance for each model considering the top-decile lift. The top-decile lift indicates how much better our top-10% prediction of churners is, compared to a 10% random prediction of churners. For the top-decile lift, we select the top-10% observations with the highest instantaneous probability to churn from the ordered churn list. Then, we determine the number of actual churners in this top-10% and divide this amount by the number of observations in top-10%. Finally, we divide this churn percentage by the actual churn percentage of the whole dataset. By repeating the above-described procedure over all deciles, we obtain the decile lift, which indicates how much better the proposed models perform compared to the benchmark model.

Table 9: Top-decile lift for the proposed models

	Top-decile lift
Churn risk	2.705
Controllable risk	2.102
Uncontrollable risk	3.764
Unknown risk	1.754

We compare the top-decile lifts in Table 9 with the top-decile lift of the benchmark model, which is by definition equal to 1. It can be seen that all top-decile lifts are higher than the top-decile lift of the benchmark model. We see that the top-10% of customers from the ordered churn list possess 2.7 times more churners than a random selection of churners possesses. Moreover, for the competing risks model, the top-10% of customers from the ordered churn list have at least 1.7 times more churners than a random selection of churners has. Note that the Uncontrollable risk top-decile lift is larger than both other competing risks top-decile lifts. This indicates that we are best able to predict the Uncontrollable risk from the competing risks model framework, as we have noticed previously.

In Figure 3 we plotted the decile lifts of the benchmark model and the churn models. The decile lift of the benchmark model is given by the dotted line. Since each randomly selected 10% of customers contains 10% of the actual churners, this line is a 45° line. Because the decile lift plots of the proposed models lie all above the decile lift plot of the benchmark model, we can conclude again that our proposed models perform better than the benchmark model. We see in Figure 3 that the top-decile lift of customers most likely to churn contains 25% of the actual churners. This implies that the call center could contact only 10% of customers, while reaching 25% of the actual churners. We also observe that the top-decile lift of customers most likely to churn due to the Uncontrollable risk, even contains 40% of the actual churners who churned due to the Uncontrollable risk. Lastly, Figure 3 demonstrates that the area between the decile lift plot of the

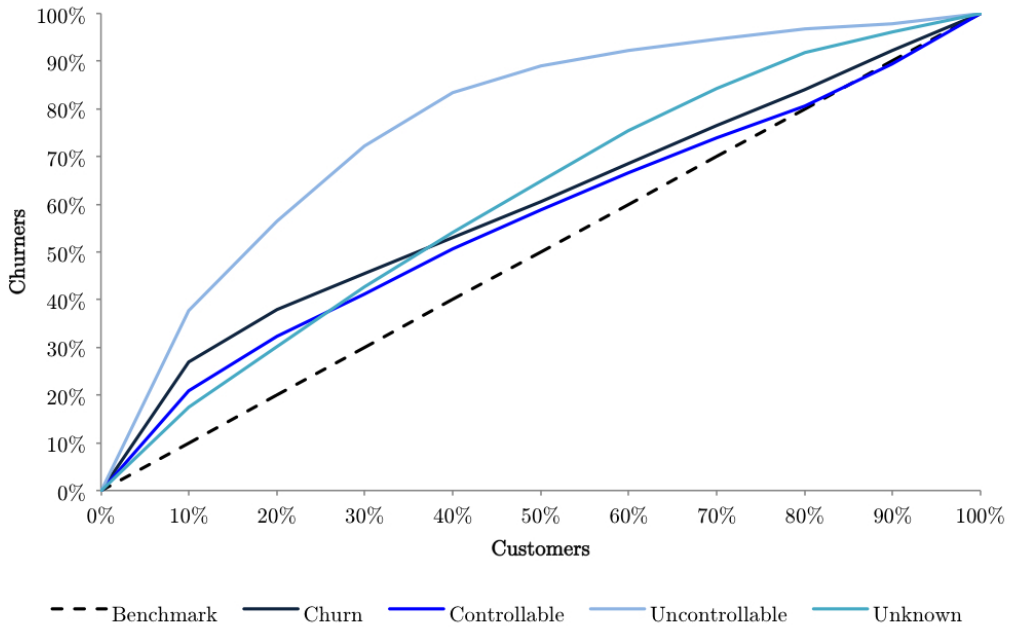


Figure 3: Decile lift plot for the churn models

Uncontrollable risk and the decile lift plot of the benchmark model is the largest of all surfaces between the churn models and the benchmark model. This again implies that the Uncontrollable risk can be modeled best over all risks.

5.4. Parameter Estimates

For the parameter estimates of the duration model, it holds that whenever a parameter is positive (negative), the hazard rate and thus the instantaneous probability to churn increases (decreases) when the value of the corresponding explanatory variable increases. This also implies that the duration time of a customer decreases (increases) when the value of the explanatory variable increases. On all occasions, we compare an increase of the instantaneous probability to churn on time t with the baseline instantaneous probability to churn on time t. Since we only incorporated categorical variables where each category of a certain categorical variable has assigned a 0 or a 1, this baseline instantaneous probability to churn corresponds to the instantaneous probability to churn of a customer whose explanatory variables equal 0 for all time periods.

The interpretation of the individual effects of a specific risk on the instantaneous probability to churn in the competing risks model is similar to the interpretation of the parameter estimates of the duration model. [59] discusses that if the effect of a certain explanatory variable on one risk is larger than for the other risks, the effect of this explanatory variable on the instantaneous probability to churn due to this risk is positive. This also indicates that the smallest parameter estimate of a certain explanatory variable over all

risks from the competing risks model shows a negative effect of this explanatory variable on the instantaneous probability to churn, due to the corresponding risk.

The parameter estimates for the duration model and the competing risks model are shown in Table 10. In this Table the parameter estimates and the associated standard error are specified, where the latter is given between parentheses. If the standard error is more than twice as small as the absolute value of the parameter estimate, the parameter estimate differs significantly from zero. For the competing risks parameter estimates, the parameters in bold are the highest parameter estimates per variable, over the three competing risks. Note that for each categorical variable, one of the categories is not incorporated into the proposed models to avoid multicollinearity. Additionally, in the Table A.13 of the Appendix, we provide a table with the (Pearson) correlations among (a subset of) our variables. The variables with the mean of absolute correlations of less than 0.05 (which implies almost no correlation) were omitted due to the space capacity. It can be clearly seen that, for the majority of the variables, there is no high enough correlation with the other ones.

Table 10 Parameter estimates

		Risk		
	Churn	Controllable	Uncontrollable	Unknown
A. Customer characteristics				
City	-0.049 (0.005)	-0.027 (0.011)	**0.105 (0.016)**	-0.041 (0.009)
Fiber	-0.792 (0.013)	-0.051 (0.025)	**1.072 (0.041)**	-0.868 (0.026)
Mosaic group				
Aged simplicity	-0.052 (0.011)	0.039 (0.027)	**0.268 (0.035)**	-0.176 (0.023)
Childs and career	-0.037 (0.011)	**0.228 (0.023)**	-0.738 (0.060)	-0.188 (0.022)
Deserved joy	-0.249 (0.012)	**0.042 (0.026)**	-0.501 (0.062)	-0.404 (0.026)
Elitist upper class	-0.236 (0.017)	**-0.003 (0.036)**	-0.419 (0.091)	-0.250 (0.032)
Freedom and space	-0.141 (0.011)	**0.052 (0.025)**	-0.672 (0.055)	-0.260 (0.023)
Golden border	-0.273 (0.014)	**-0.035 (0.029)**	-0.942 (0.091)	-0.370 (0.029)
Good city life	0.000 (0.010)	-0.113 (0.027)	-0.110 (0.038)	**0.101 (0.018)**
Mature middle class	-0.008 (0.011)	**0.120 (0.024)**	-0.116 (0.039)	-0.070 (0.032)
Modal households	-0.018 (0.011)	**0.297 (0.023)**	-0.510 (0.047)	-0.111 (0.021)
Rural life	-0.360 (0.015)	**-0.297 (0.034)**	-0.454 (0.068)	-0.362 (0.029)
Starting together	0.092 (0.010)	**0.273 (0.023)**	-0.023 (0.029)	-0.004 (0.019)
Urban balancers	0.027 (0.011)	-0.507 (0.038)	**0.226 (0.029)**	-0.041 (0.021)
Young digitals	0.094 (0.009)	0.031 (0.023)	**0.176 (0.026)**	0.106 (0.016)
B. Product characteristics				
Product type				
aTV	-0.269 (0.043)	**-0.184 (0.074)**	-0.678 (0.162)	-0.819 (0.126)
Internet	-0.729 (0.014)	-1.215 (0.030)	-0.801 (0.063)	**-0.429 (0.028)**
Internet, phone	-0.403 (0.010)	**-0.445 (0.022)**	-0.621 (0.046)	-0.457 (0.023)
Internet, iTV	0.016 (0.008)	-0.301 (0.020)	-0.117 (0.024)	**0.323 (0.015)**
Internet, aTV, iTV	0.347 (0.027)	**0.360 (0.042)**	0.136 (0.087)	0.017 (0.072)
Phone, aTV	-0.504 (0.053)	-0.448 (0.088)	**-0.196 (0.144)**	-1.395 (0.169)

Table 10 – continued from previous page

	Risk			
	Churn	Controllable	Uncontrollable	Unknown
Internet, phone, aTV	-0.427 (0.023)	-0.519 (0.037)	**-0.453 (0.088)**	-0.886 (0.070)
Purchase channel				
Door-to-door	0.121 (0.019)	-0.274 (0.037)	**0.072 (0.067)**	-0.052 (0.047)
Online	-0.078 (0.009)	-0.065 (0.023)	**0.377 (0.021)**	-0.091 (0.016)
Phone	-0.053 (0.010)	-0.103 (0.025)	**0.220 (0.026)**	-0.042 (0.019)
Resell	-0.164 (0.012)	-0.122 (0.029)	**0.236 (0.034)**	-0.163 (0.022)
Unknown	-0.011 (0.008)	0.344 (0.021)	**0.635 (0.034)**	-0.040 (0.016)
Price				
< €20	0.721 (0.015)	0.598 (0.034)	**0.772 (0.071)**	0.677 (0.029)
€20 - €30	0.491 (0.011)	0.397 (0.023)	0.404 (0.057)	**0.534 (0.024)**
€40 - €50	0.162 (0.011)	**0.376 (0.022)**	0.168 (0.032)	0.154 (0.023)
€50 - €60	0.229 (0.011)	**0.283 (0.025)**	0.049 (0.028)	0.121 (0.023)
> €60	0.666 (0.025)	0.501 (0.047)	**0.577 (0.070)**	0.435 (0.054)
VAS				
Entertainment	1.126 (0.141)	**0.981 (0.507)**	0.669 (0.275)	0.939 (0.256)
Eredivisie	0.062 (0.021)	-0.173 (0.065)	-0.044 (0.052)	**0.034 (0.043)**
Erotica	0.308 (0.109)	0.285 (0.189)	**0.986 (0.182)**	-0.002 (0.268)
Film1	0.375 (0.023)	-0.268 (0.080)	**0.230 (0.043)**	-0.259 (0.056)
FOX	0.015 (0.026)	0.073 (0.045)	**0.315 (0.067)**	0.140 (0.056)
German	-0.634 (0.075)	-0.376 (0.148)	**0.207 (0.137)**	-0.380 (0.136)
HBO	-0.115 (0.015)	-0.380 (0.043)	**0.153 (0.034)**	-0.105 (0.031)
HD	1.598 (0.044)	**1.238 (0.057)**	0.070 (0.183)	1.149 (0.086)
Hindi	-0.164 (0.055)	-0.195 (0.148)	**-0.059 (0.106)**	-0.386 (0.115)
Kids	0.163 (0.020)	-0.068 (0.061)	0.152 (0.041)	**0.231 (0.042)**
Nature	0.714 (0.173)	0.358 (0.460)	0.257 (0.321)	**0.789 (0.270)**
PC security	-0.186 (0.012)	-0.403 (0.032)	-0.327 (0.041)	**-0.126 (0.023)**
Plus	0.049 (0.008)	-0.234 (0.020)	**0.175 (0.022)**	0.133 (0.016)
Premium	0.434 (0.041)	-0.096 (0.152)	**0.391 (0.077)**	0.278 (0.091)
Sport1	0.262 (0.026)	0.152 (0.072)	**0.271 (0.054)**	0.213 (0.054)
Turkish	-0.068 (0.051)	**0.024 (0.144)**	0.020 (0.103)	-0.264 (0.108)
Videoland	0.613 (0.021)	0.282 (0.066)	**0.521 (0.039)**	0.381 (0.045)
C. Billing information				
WP	0.106 (0.013)	0.015 (0.032)	-0.368 (0.086)	**0.280 (0.022)**
WPSD	1.006 (0.014)	-0.058 (0.041)	**3.754 (0.084)**	0.125 (0.026)
D. Customer service information				
Number of calls t-1				
1-2	-	-	**1.255 (0.012)**	-
3-5	-	-	**1.272 (0.019)**	-
6-15	-	-	**1.297 (0.029)**	-
16-30	-	-	**1.199 (0.087)**	-
> 31	-	-	**1.163 (0.203)**	-
Topic t-1				
Administration	1.305 (0.018)	**1.245 (0.015)**	-	0.391 (0.011)

Table 10 – continued from previous page

	Risk			
	Churn	Controllable	Uncontrollable	Unknown
Usage	-0.289 (0.015)	**0.181 (0.014)**	-	0.082 (0.011)
Payment	-0.212 (0.018)	-0.005 (0.019)	-	**-0.001 (0.013)**
Product	0.922 (0.015)	0.675 (0.014)	-	**1.788 (0.009)**
Question	0.826 (0.016)	0.394 (0.014)	-	**0.875 (0.010)**
Status order	-0.005 (0.019)	0.052 (0.016)	-	**0.139 (0.012)**
E. Failure information				
Number of failures t-1				
1-2	0.298 (0.009)	1.024 (0.011)	0.888 (0.019)	**1.181 (0.008)**
3-5	0.225 (0.017)	0.998 (0.018)	0.815 (0.028)	**1.207 (0.013)**
6-15	0.139 (0.020)	0.616 (0.040)	0.838 (0.045)	**1.150 (0.022)**
> 15	0.206 (0.075)	0.231 (0.211)	-0.024 (0.112)	**0.675 (0.047)**
F. Online activities				
Number of activities TSP t-1				
1-50	0.076 (0.034)	**0.402 (0.036)**	-0.341 (0.087)	0.094 (0.032)
> 51	-0.243 (0.058)	**0.249 (0.056)**	-0.845 (0.163)	-0.338 (0.062)
Number of activities fTSP t-1				
1-10	0.355 (0.038)	0.317 (0.037)	**0.359 (0.084)**	0.300 (0.035)
11-30	0.945 (0.057)	0.521 (0.059)	0.373 (0.114)	**0.687 (0.057)**
31-50	1.262 (0.089)	0.695 (0.092)	0.592 (0.188)	**1.027 (0.085)**
51-70	1.551 (0.122)	0.665 (0.135)	0.608 (0.238)	**1.215 (0.115)**
71-200	1.679 (0.104)	0.907 (0.128)	0.821 (0.262)	**1.365 (0.112)**
> 201	1.302 (0.311)	0.345 (0.457)	0.792 (0.598)	**0.993 (0.335)**

5.4.1. Duration Model

The parameter estimates of the duration model are given in the Churn risk column of Table 10. We discarded the mosaic group *social housing* from the proposed models, to avoid multicollinearity. We see that the parameter estimates of the mosaic groups *starting together, urban balancers* and *young digitals* are all positive. These mosaic groups mainly consist of young customers of the TSP. The positive parameter estimates indicate that young customers are more likely to churn than customers from all other mosaic groups. Note that parameter estimates of the mosaic group variables all differ significantly from zero.

For the product type variables, we discarded the product *Internet, phone, iTV* to avoid multicollinearity. We see that the smallest parameter estimate corresponds to the product *Internet*, which indicates that the instantaneous probability to churn is the smallest for customers with the product Internet, compared to the instantaneous probability to churn of customers with another product. Since the TSP assumes that customers who own more products are less likely to churn, the parameter estimate of the variable *Internet*

is unexpected. A potential reason is that customers who own the product Internet, cannot switch to other providers, because other providers no longer offer the product Internet. Note that all parameter estimates of the product type variables differ significantly from zero.

The variable *shop* is discarded within the purchase channel variables. We see that the *door-to-door* purchase channel has the largest parameter estimate of all purchase channel parameters. This suggests that customers who purchased their products via the door-to-door purchase channel have a higher instantaneous probability to churn than customers who purchased their products via other purchase channels. This is in line with our expectations since products purchased via the door-to-door channel are usually purchased due to too much pressure of the salesman. Note that all parameter estimates of the purchase channel variables differ significantly from zero.

Table 10 also shows that the parameter estimates of the price variables are all positive. This indicates that each price different from a price between €30 and €40 (the baseline price, which is discarded from the proposed models), induces a higher instantaneous probability to churn. Note that all parameter estimates of the price variables differ significantly from zero.

For the VAS variables, especially the parameter estimate for the VAS *HD* is large, which implies that customers who purchase this VAS are very likely to churn, compared to customers who do not purchase the VAS *HD*. A likely reason is that new customers receive the VAS *HD* for free, while customers who are purchasing this VAS for more than three years have to pay. This induces dissatisfaction concerning the TSP among customers who still have to pay for the HD VAS.

The parameter estimates of the LDA generated topics as variables are also given in Table 10. These variables are not constant and depend on time $t-1$. It can be seen that if a customer calls the customer service center on time $t-1$ about a *product*, a *question* in general, or an *administrative* problem the parameter estimate is positive. Customers who call the customer service center about these topics on time $t-1$, are more likely to churn than customers who do not call the customer service center about these topics on time $t-1$. The parameter estimates of calls about *usage*, the *order status* and *payments* are all negative, indicating a decrease in the instantaneous probability to churn when a customer calls about these topics. All parameter estimates of the topic variables differ significantly from zero.

For both the number of failures on time $t-1$ and the number of activities on the website of the fTSP on time $t-1$, the parameter estimates are positive. This implies an increase in the instantaneous probability to churn when customers eventually perceive a failure, or visited a webpage of the TSP its family companies, compared to the baseline instantaneous probability to churn. This is in line with our expectations. Observe that the number of failures and the number of activities is not constant, and differ over time. Moreover, note that all parameter estimates of the failure information and the online activities differ significantly from zero.

5.4.2. Competing Risks Model

For the competing risks model, the parameter estimates are given in the Controllable risk column, the Uncontrollable risk column and the Unknown risk column of Table 10. At first sight, we see that the differences in parameter estimates between the three risks can be large. Remember that all results only hold when we assume the *ceteris paribus* principle.

As mentioned before, the parameter estimates in bold font are the largest parameter estimates for that variable, among the three risks. It can be seen that the parameter estimate for having a *fiber* connection is the largest for the Uncontrollable risk. Therefore, for customers with a fiber connection the instantaneous probability to churn due to the Uncontrollable risk is larger than the instantaneous probability to churn due to the other risks, whilst all other variables are equal. This result is unexpected since the TSP assumes that the connection type is irrelevant for the reason for which a customer churns.

All parameter estimates for the purchase channel variables are the largest for the Uncontrollable risk. This implies that customers who do not purchase their products or services in a shop are more likely to churn due to the Uncontrollable risk than due to any other risk.

The table also shows that for all risks, a price between €30 and €40 would be optimal for reducing the instantaneous probability to churn in general. Since the parameter estimates for the price variables *<€20* and *>€60* are the largest for the Uncontrollable risk, customers with a relatively cheap or relatively expensive contract have a higher instantaneous probability to churn due to the Uncontrollable risk, than due to any other risk. Similarly, the parameter estimates for the price variables *€40 - €50* and *€50 - €60* are the largest for the Controllable risk, which implies that customers with contracts in these price classes are more likely to churn due to the Controllable risk than due to any other risk.

It can be seen that the parameter estimates of VAS variables are often the largest for the Uncontrollable risk. This shows that most of the VASs induce churn due to the Uncontrollable risk.

The parameter estimate of the *WPSD* variable is the largest for Uncontrollable risk, among the three risks. This suggests that customers who have seen a warning page and who have had a soft disconnect are more likely to churn due to the Uncontrollable risk, than due to any other risk. As the delay in payments is part of the Uncontrollable risk, this observation is in line with our expectations.

For the Uncontrollable risk, we incorporated the number of calls between the customer and the customer service center on time $t-1$ in the competing risks model, instead of the topic variables on time $t-1$, since this model turned out to perform best, as discussed in Section 5.2. For these variables, all parameter estimates are positive, which implies an increase in the instantaneous probability to churn when a customer calls the customer service center, irrespectively of the number of calls. The parameter estimates of the

Controllable risk for the topics *usage* and *administration* are larger than for the Unknown risk. This implies that a call between a customer and the customer service center about *usage* and *administration* increases the instantaneous probability to churn due to the Controllable risk, which is in line with our expectations. Furthermore, the parameter estimates of calls about *payments*, *products*, *questions in general* and the *status of the order* are larger than for the Controllable risk. This indicates that a call about one of these risks increases the instantaneous probability of churn due to the Unknown risk.

All parameter estimates for the failure variables are the largest for the Unknown risk, indicating that a failure on time $t-1$ induces a higher instantaneous probability to churn due to the Unknown risk. In general, the parameter estimates for the failure variables are positive, which implies that each occurred failure increases the probability of churn.

The parameter estimates for the number of online activities on the website of the TSP on time $t-1$ are the largest for the Controllable risk. This suggests that activity on the website of the TSP on time $t-1$ increases the likeliness to churn due to the Controllable risk. Probably, customers who churn due to the Controllable risk often visit the TSP's website to compare their current contract to offers of the TSP's competitors. For the online activities on the website of a family company of the TSP, the parameter estimates are in most cases the largest for the Unknown risk. Thereby, all parameter estimates for the fTSP variables are positive, indicating that the probability to churn due to any risk increases when a customer visits a webpage of an fTSP. This follows our expectations, as visiting a competitor's website means searching for another offer.

6. Conclusion

6.1. Practical and Theoretical Contributions

The main practical contribution of our work is that we investigated the applicability of a competing risks model and LDA to predict the reasons for which customers churn. On the theoretical side, while LDA is often applied for gathering topics from documents, it has never been applied to conversations between customers and the customer service center to incorporate their topics as variables into the competing risks model. We have shown that the data empirically supports the provided theory and customer service data provides useful signals for identifying churn and its reasons.

Reducing the churn rate is very relevant in marketing since the costs of acquiring a new customer are five times higher than maintaining an existing one. Knowledge of the propensity to churn and its potential reason for each customer provides the opportunity to create a personalised strategy for each customer and increase retention. One of the questions we aimed to answer is whether a competing risks model with incorporated customer service data as variables performs better than a competing risks model without it.

For predicting the propensity to churn, we have proposed two duration models: one model without topic variables, and one model with topic variables. The topic variables have been created by employing an LDA model on the textual customer service center data. We also aggregated the reasons for which customers could churn in three competing risks: the Controllable risk, the Uncontrollable risk, and the Unknown risk. Thereafter, we have created a competing risks model that could simultaneously model churn and the reason for which a customer churns. Again, we made one competing risks models without topic variables, and one competing risks model with topic variables.

All models for predicting churn, and the ones for simultaneously predicting the reason for which a customer churns, outperform the benchmark model. The in-sample performance measures have shown that the duration model with topic variables outperforms the duration model without topic variables. This indicates that we are more successful in modeling the propensity to churn, by including the topics of the calls between a customer and the customer service center of the TSP. We have also seen that for the Controllable risk, and for the Unknown risk, the models with topic variables outperform the models without topic variables. This implies that taking the topic of a call between a customer and the TSP's customer service center into account improves the model for the Controllable risk and the Unknown risk. Nonetheless, for the Uncontrollable risk, we have seen that the model without topic variables outperforms the model with topic variables.

We conclude that the effects of the incorporated variables on the propensity to churn differ significantly between the various risks. This implies that the behavior of customers who churn due to varying risks differs. Focusing on the topic variables, we have seen that it does not necessarily mean that the customer's propensity to churn increases when the customer has had contact with the TSP's customer service center. For some topics the propensity to churn even decreases. Other strong effects on the propensity to churn are the occurrences of failures, which increases the propensity to churn due to the Unknown risk. Moreover, the probability to churn due to the Controllable risk of customers who visit online webpages of the TSP increases, while the probability to churn due to the Unknown risk of customers who visit online webpages of a family brand of the TSP increases. We have also seen that the price of the contract is optimal between €30 and €40.

Using the developed churn models provides business managers with an opportunity to act proactively rather than reactively while doing it with a comparatively high level of accuracy. By this, we mean that they would be able to predict the customers' churn period and reason long enough in advance. Combining it with the previously discussed fact that attracting new customers is very costly under the conditions of fierce competition in the telecom industry, we can see a clear managerial application. Companies would be able to design personalised offers to potential churners. These offers would specifically address the reason for which a customer is likely to leave. This, in turn, would advance the efforts of keeping the customer satisfied. Additionally, by utilising the knowledge extracted from the churn model, managers in the companies will be

able to improve the business decision making processes (e.g., targeted marketing - knowing the right audience) by making these more data-driven.

6.2. Limitations and Future Work

Nevertheless, our work has a few limitations. In particular, we can make several improvements in the approach of the churn models. One drawback of our churn models is that we only incorporate time-varying variables at time $t-1$ in a model that models the instantaneous probability to churn at time t. An improvement would be to predict the unknown time-varying explanatory variables at time t, when predicting the probability to churn at time t. In this case, it would be possible to incorporate current predicted behavior in the model instead of just past behavior. In order to predict the behavior of a customer, we should set up a predictive model that predicts the number of calls, the topics of the calls, the number of failures, and the number of online activities, all at time t.

In future research, it would be interesting to predict the reason for which a customer churns more specifically. For example, the Controllable risk can be divided into several more specific risks, such as churn due to bad services, churn due to a better offer of a competitor, or churn due to a perceived high price of the contract. The Uncontrollable risk can also be divided into, for example, two risks, churn due to migration, and churn due to death. When we are able to predict the reason for which a customer churns more specifically, the TSP's offers can be more personalized, which makes churn even easier to prevent.

References

[1] Internetstats, 2020. URL: http://www.internetworldstats.com/stats.htm.

[2] A. M. Hughes, 2020. URL: http://www.dbmarketing.com/telecom/churnreduction.html.

[3] C. Grönroos, A service quality model and its marketing implications, European Journal of Marketing 18 (1984) 36–44.

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

[5] K. Coussement, D. Van den Poel, Integrating the voice of customers through call center emails into a decision support system for churn prediction, Information and Management 45 (2008) 164–174.

[6] B. Huang, M. Kechadi, B. Buckley, Customer churn prediction in telecommunications, Expert Systems with Applications 39 (2012) 1414–1425.

[7] M. Braun, D. Schweidel, Modeling customer lifetimes with multiple causes of churn, Marketing Science 30 (2011) 881–902.

[8] P. Fader, B. Hardie, Customer-base valuation in a contractual setting: The perils of ignoring heterogeneity, Marketing Science 29 (2010) 85–93.

[9] W. Reinartz, V. Kumar, The impact of customer relationship characteristics on profitable lifetime duration, Journal of Marketing 67 (2003) 77–99.

[10] X. Xu, J. Y. Thong, V. Venkatesh, Effects of ict service innovation and complementary strategies on brand equity and customer loyalty in a consumer technology market, Information Systems Research 25 (2014) 710–729.

[11] S. Gupta, D. Lehmann, J. Stuart, Valuing customers, Journal of Marketing Research 41 (2004) 7–18.

[12] Y. Kim, Toward a successful CRM: variable selection, sampling, and ensemble, Decision Support Systems 41 (2006) 542–553.

[13] P.-Y. Chen, L. M. Hitt, Measuring switching costs and the determinants of customer retention in Internet-enabled businesses: A study of the online brokerage industry, Information Systems Research 13 (2002) 255–274.

[14] T. Dierkes, M. Bichler, R. Krishnan, Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with Markov logic networks, Decision Support Systems 51 (2011) 361–371.

[15] S. Keaveney, Customer switching behavior in service industries: An exploratory study, Journal of Marketing (1995) 71–82.

[16] R. Rust, A. Zahorik, T. Keiningham, Return on quality (ROQ): Making service quality financially accountable, Journal of Marketing (1995) 58–70.

[17] S. Hung, D. Yen, H. Wang, Applying data mining to telecom churn management, Expert Systems with Applications 31 (2006) 515–524.

[18] J. Burez, D. Van den Poel, Crm at a pay-tv company: Using analytical models to reduce customer attrition by targeted marketing for subscription services, Expert Systems with Applications 32 (2007) 277–288.

[19] J. Burez, D. Van den Poel, Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department, Expert Systems with Applications 35 (2008) 497–514.

[20] S. Qureshi, A. Rehman, A. Qamar, A. Kamal, A. Rehman, Telecommunication subscribers' churn prediction model using machine learning, in: Proceedings of the 8th International Conference on Digital Information Management, IEEE, 2013, pp. 131–136.

[21] C. Wei, I. Chiu, Turning telecommunications call details to churn prediction: A data mining approach, Expert Systems with Applications 23 (2002) 103–112.

[22] K. Coussement, D. Van den Poel, Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques, Expert Systems with Applications 34 (2008) 313–327.

[23] G. Xia, W. Jin, Model of customer churn prediction on support vector machine, Systems Engineering-Theory and Practice 28 (2008) 71–77.

[24] M. Mozer, R. Wolniewicz, D. Grimes, E. Johnson, H. Kaushansky, Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry, IEEE Transactions on Neural Networks 11 (2000) 690–696.

[25] H. Hwang, T. Jung, E. Suh, An ltv model and customer segmentation based on customer value: A case study on the wireless telecommunication industry, Expert Systems with Applications 26 (2004) 181–188.

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

[27] E. Lima, C. Mues, B. Baesans, Domain knowledge integration in data mining using decision tables: Case studies in churn prediction, Journal of the Operational Research Society 60 (2009) 1096–1106.

[28] W. Verbeke, D. Martens, C. Mues, B. Baesans, Building comprehensible customer churn prediction models with advanced rule induction techniques, Expert Systems with Applications 38 (2011) 2354–2364.

[29] M. Farquad, V. Ravi, S. Raju, Churn prediction using comprehensible support vector machine: An analytical crm application, Applied Soft Computing 19 (2014) 31–40.

[30] B. Larivière, D. Van den Poel, Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services, Expert Systems with Applications 27 (2004) 277–285.

[31] Z. Jamal, R. Bucklin, Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach, Journal of Interactive Marketing 20 (2006) 16–29.

[32] A. Cameron, P. Trivedi, Microeconometrics, Cambridge University Press, 2005.

[33] D. Cox, Regression models and life-tables, Journal of the Royal Statistical Society. Series B (Methodological) (1972) 187–220.

[34] M. Burda, M. Harding, J. Hausman, A bayesian semiparametric competing risks model with unobserved heterogeneity, Journal of Applied Econometrics 30 (2014) 353–376.

[35] J. Beyersmann, A. Allignol, M. Schumacher, Competing Risks and Multistate Models with R, Springer Science and Business Media, 2011.

[36] M. Pintilie, Competing Risks: A Practical Perspective, John Wiley and Sons, 2006.

[37] A. Han, J. Hausman, Flexible parametric estimation of duration and competing risks models, Journal of Applied Econometrics 5 (1990) 1–28.

[38] B. McCall, Unemployment insurance rules, joblessness, and part-time work, Econometrica: Journal of the Econometric Society (1996) 647–682.

[39] N. Vilcassim, D. Jain, Modeling purchase-timing and brand-switching behavior incorporating explanatory variables and unobserved heterogeneity, Journal of Marketing Research (1991) 29–41.

[40] P. Chintagunta, Inertia and variety seeking in a model of brand-purchase timing, Marketing Science 17 (1998) 253–270.

[41] R. Srinivasan, G. Lilien, A. Rangaswamy, Survival of high tech firms: The effects of diversity of product-market portfolios, patents, and trademarks, International Journal of Research in Marketing 25 (2008) 119–128.

[42] E. de Haan, E. Menichelli, The Incremental Value of Unstructured Data in Predicting Customer Churn, Technical Report 20-105, Marketing Science Institute, 2020.

[43] D. Thorleuchter, D. Van den Poel, Predicting e-commerce company succes by mining the text of its publicly-accessible website, Expert Systems with Applications 39 (2012) 13026–13034.

[44] A. De Caigny, K. Coussement, K. W. De Bock, S. Lessmann, Incorporating textual information in customer churn prediction models based on a convolutional neural network, International Journal of Forecasting 36 (2019) 1563–1578.

[45] D. M. Blei, J. D. Lafferty, Topic models, in: A. Srivastava, M. Sahami (Eds.), Text mining: classification, clustering, and applications, Chapman & Hall/CRC, 2009, pp. 71–94.

[46] D. Blei, J. Lafferty, A correlated topic model of science, Annals of Applied Statistics (2007) 17–35.

[47] T. K. Landauer, P. W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Processes 25 (1998) 259–284.

[48] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[49] J. Wang, M. Zaki, H. Toivonen, D. Shasha, Introduction to Data Mining in Bioinformatics, Springer London, 2005.

[50] J. Wang, S. Faridani, P. Ipeirotis, Estimating the completion time of crowdsourcesd tasks using survival analysis models, in: Proceeding of the 2011 Crowdsourcing for Search and Data Mining Conference, 2011, pp. 31–38.

[51] S. Gerrish, D. Blei, Predicting legislative roll calls from text, in: Proceedings of the 28th International Conference on Machine Learning, Omnipress, 2011, pp. 489–496.

[52] T. Yano, N. Smith, J. Wilkerson, Textual predictors of bill survival in congressional committees, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, 2012, pp. 793–802.

[53] J. Grimmer, B. Stewart, Text as data: The promise and pitfalls of automatic content analysis methods for political texts, Political Analysis 21 (2013) 267–297.

[54] Y. Wang, R. Kraut, J. Levine, To stay or to leave? The relationship of emotional and informational support to commitment in online health support groups, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, ACM, 2012, pp. 833–842.

[55] X. Wang, K. Zhao, N. Street, Social support and user engagement in online health communities, in: Proceedings of the 2014 International Conference on Smart Health, Springer, 2014, pp. 97–110.

[56] A. Nikfarjam, E. Emadzadeh, S. Muthaiyah, Text mining approaches for stock market prediction, in: Proceedings of the 2nd International Conference on Computer and Automation Engineering, volume 4, IEEE, 2010, pp. 256–260.

[57] A. De Caigny, K. Coussement, K. W. De Bock, S. Lessmann, Incorporating textual information in customer churn prediction models based on a convolutional neural network, International Journal of Forecasting 36 (2020) 1563–1578.

[58] W. Darling, A theoretical and practical implementation tutorial on topic modeling and gibbs sampling, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011) 642–647.

[59] J. Thomas, On the interpretation of covariate estimates in independent competing risks models, Bulletin of Economic Research 48 (1996) 27–39.

Appendix A. Variables

In this appendix we list all existing categories of some of the categorical variables explained in Section 3 and provide the pairwise (Pearson) correlations between variables.

Categorical Variables

Mosaic Group: Table A.11 shows the existing categories for the *mosaic group* variable. This table also shows the description of each category. We denote the size of the household with *size hh.*

Table A.11: Mosaic group categories

Category	Age	Size hh.	Income	Education level	Housing
Young digitals	<30	1-2	low	low or high	rent
Urban balancers	<40	1-2	low	low or high	rent or bought
Starting together	25-45	1-4	low	low	rent
Good city life	25-60	1-2	low or high	high	bought
Modal households	25-55	>1	low - high	average	bought
Childs and career	25-55	>2	average or high	average or high	bought
Social housing	45-65	1-3	low	low	rent
Mature middle class	45-75	1-4	low or average	low - high	bought
Freedom and space	35-65	>1	average or high	average	bought
Golden border	45-75	>1	average or high	high	bought
Elitist upper class	>45	>0	high	high	bought
Rural life	45-75	>1	average or high	low - high	bought
Deserved joy	>55	1-2	average or high	average or high	bought
Aged simplicity	>65	1-2	low	low	rent

Play Name: The existing categories for the play name variable are: *aTV*, *Internet*, *Internet* and *phone*, *Internet* and *iTV*, *Internet* and *aTV* and *iTV*, *phone* and *aTV*, *Internet* and *phone* and *aTV*, *Internet* and *phone* and *iTV*.

Channel: The existing categories for the channel variable are: *online*, *phone*, *resell*, *door-to-door*, *shop*, and *unknown*.

Value Added Service: Table A.12 shows the existing categories for the *VAS* variable.

Correlations

Table A.13 shows the pairwise Pearson correlations between (a subset of) variables. For space reasons, variable which had the mean of the absolute values of the correlations less than 0.05 were omitted.

Table A.12: VAS categories

Category	Description
Eredivisie	A TV channel that mainly broadcasts the Dutch soccer league.
Erotica	A TV channel that only broadcasts erotica.
Film1	A TV channel that mainly broadcasts movies and series.
HBO	A TV channel that mainly broadcasts movies and series.
HD	An extra service for having brighter views.
Hindi	A TV channel that mainly broadcasts Indian programs.
Kids	A TV channel that only broadcasts entertainment for kids.
Nature	A TV channel that mainly broadcasts documentaries.
PC security	An extra service that secures your PC.
Plus	A TV channel that mainly broadcasts movies, series and extra documentaries.
Premium	A TV channel that mainly broadcasts movies, series and extra documentaries.
Sport1	A TV channel that mainly broadcasts sports.
Turkish	A TV channel that mainly broadcasts Turkish programs.
Videoland	A TV channel that mainly broadcasts movies and series.
German	A TV channel that mainly broadcasts German programs.
Entertainment	A TV channel that only broadcasts erotica.

Table A.13: Pearson Correlations. Part 1

	Time	Status	#Fails	#Contacts	Fiber	City	Offer1	Offer2	Offer3	Offer4	WP	WPSD	WPSDHD
Time	1.000	-0.060	0.088	-0.008	0.035	0.006	0.038	0.107	0.112	-0.123	-0.036	-0.035	-0.062
Status	-0.060	1.000	-0.079	0.012	-0.048	-0.005	0.007	0.056	0.026	-0.022	0.072	0.093	0.186
#Fails	0.088	-0.079	1.000	-0.062	-0.132	0.283	-0.013	0.102	0.008	0.030	0.001	-0.003	-0.017
#Contacts	-0.008	0.012	-0.062	1.000	0.125	-0.034	0.001	-0.070	-0.034	-0.026	0.030	0.028	0.004
Fiber	0.035	-0.048	-0.132	0.125	1.000	-0.149	0.096	-0.264	-0.122	-0.265	-0.149	-0.137	-0.068
City	0.006	-0.005	0.283	-0.034	-0.149	1.000	-0.026	0.100	-0.004	0.070	0.028	0.025	0.014
Offer1	0.038	0.007	-0.013	0.001	0.096	-0.026	1.000	-0.025	-0.018	-0.027	-0.014	-0.013	-0.007
Offer2	0.107	0.056	0.102	-0.070	-0.264	0.100	-0.025	1.000	-0.183	-0.275	0.016	0.013	0.004
Offer3	0.112	0.026	0.008	-0.034	-0.122	-0.004	-0.018	-0.183	1.000	-0.193	-0.032	-0.026	-0.019
Offer4	-0.123	-0.022	0.030	-0.026	-0.265	0.070	-0.027	-0.275	-0.193	1.000	0.032	0.024	0.007
WP	-0.036	0.072	0.001	0.030	-0.149	0.028	-0.014	0.016	-0.032	0.032	1.000	0.782	0.385
WPSD	-0.035	0.093	-0.003	0.028	-0.137	0.025	-0.013	0.013	-0.026	0.024	0.782	1.000	0.492
WPSDHD	-0.062	0.186	-0.017	0.004	-0.068	0.014	-0.007	0.004	-0.019	0.007	0.385	0.492	1.000
P<20	-0.021	0.024	0.031	-0.030	-0.084	0.017	0.229	0.399	-0.077	-0.118	-0.009	-0.008	-0.001
P20-30	0.111	0.047	0.087	-0.068	-0.270	0.076	-0.028	0.668	0.312	-0.298	-0.002	-0.001	-0.008
P30-40	-0.052	-0.010	0.021	-0.041	-0.434	0.045	-0.042	-0.426	-0.100	0.627	0.050	0.043	0.021
P40-50	-0.113	-0.054	-0.088	0.104	0.621	-0.096	-0.024	-0.248	-0.070	-0.242	-0.082	-0.075	-0.043
P50-60	0.077	0.000	-0.054	0.043	0.288	-0.053	-0.017	-0.175	-0.116	-0.186	0.034	0.036	0.032
Plus	-0.035	0.000	-0.035	0.039	0.055	-0.032	-0.018	-0.180	-0.127	0.040	0.110	0.101	0.069
MaxShould	0.815	-0.003	0.013	-0.051	0.028	-0.002	0.047	0.150	0.141	-0.151	-0.048	-0.047	-0.075
Topic1	-0.030	0.005	-0.040	0.069	0.059	-0.019	-0.003	-0.035	-0.023	-0.001	0.010	0.007	0.001
Topic2	-0.035	-0.003	-0.043	0.087	0.114	-0.026	-0.004	-0.063	-0.031	-0.015	0.015	0.012	0.001
Topic3	-0.021	-0.033	-0.046	0.076	0.017	-0.022	-0.010	-0.065	-0.024	0.007	0.015	0.009	-0.011
Topic4	-0.037	0.014	-0.038	0.081	0.080	-0.027	-0.006	-0.064	-0.011	-0.025	-0.008	-0.008	-0.005
Topic5	-0.053	-0.018	-0.031	0.082	0.052	-0.015	-0.007	-0.049	-0.029	0.008	-0.011	-0.011	-0.008
Topic6	-0.052	-0.013	-0.025	0.050	-0.007	-0.002	-0.009	-0.052	-0.031	0.023	0.156	0.141	0.056

Pearson Correlations. Part 2

	P<20	P20-30	P30-40	P40-50	P50-60	Plus	MaxShould	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
Time	-0.021	0.111	-0.052	-0.113	0.077	-0.035	0.815	-0.030	-0.035	-0.021	-0.037	-0.053	-0.052
Status	0.024	0.047	-0.010	-0.054	0.000	0.000	-0.003	0.005	-0.003	-0.033	0.014	-0.018	-0.013
#Fails	0.031	0.087	0.021	-0.088	-0.054	-0.035	0.013	-0.040	-0.043	-0.046	-0.038	-0.031	-0.025
#Contacts	-0.030	-0.068	-0.041	0.104	0.043	0.039	-0.051	0.069	0.087	0.076	0.081	0.082	0.050
Fiber	-0.084	-0.270	-0.434	0.621	0.288	0.055	0.028	0.059	0.114	0.017	0.080	0.052	-0.007
City	0.017	0.076	0.045	-0.096	-0.053	-0.032	-0.002	-0.019	-0.026	-0.022	-0.027	-0.015	-0.002
Offer1	0.229	-0.028	-0.042	-0.024	-0.017	-0.018	0.047	-0.003	-0.004	-0.010	-0.006	-0.007	-0.009
Offer2	0.399	0.668	-0.426	-0.248	-0.175	-0.180	0.150	-0.035	-0.063	-0.065	-0.064	-0.049	-0.052
Offer3	-0.077	0.312	-0.100	-0.070	-0.116	-0.127	0.141	-0.023	-0.031	-0.024	-0.011	-0.029	-0.031
Offer4	-0.118	-0.298	0.627	-0.242	-0.186	0.040	-0.151	-0.001	-0.015	0.007	-0.025	0.008	0.023
WP	-0.009	-0.002	0.050	-0.082	0.034	0.110	-0.048	0.010	0.015	0.015	-0.008	-0.011	0.156
WPSD	-0.008	-0.001	0.043	-0.075	0.036	0.101	-0.047	0.007	0.012	0.009	-0.008	-0.011	0.141
WPSDHD	-0.001	-0.008	0.021	-0.043	0.032	0.069	-0.075	0.001	0.001	-0.011	-0.005	-0.008	0.056
P<20	1.000	-0.121	-0.184	-0.107	-0.075	-0.078	-0.012	-0.016	-0.027	-0.031	-0.027	-0.024	-0.026
P20-30	-0.121	1.000	-0.463	-0.269	-0.190	-0.196	0.149	-0.036	-0.065	-0.059	-0.055	-0.050	-0.055
P30-40	-0.184	-0.463	1.000	-0.408	-0.289	0.080	-0.065	-0.017	-0.037	0.018	-0.023	0.001	0.024
P40-50	-0.107	-0.269	-0.408	1.000	-0.168	0.060	-0.147	0.054	0.106	0.048	0.082	0.061	0.031
P50-60	-0.075	-0.190	-0.289	-0.168	1.000	0.107	0.082	0.017	0.027	0.012	0.024	0.006	0.015
Plus	-0.078	-0.196	0.080	0.060	0.107	1.000	-0.050	0.013	0.031	0.036	0.020	0.024	0.056
MaxShould	-0.012	0.149	-0.065	-0.147	0.082	-0.050	1.000	-0.063	-0.083	-0.081	-0.066	-0.083	-0.086
Topic1	-0.016	-0.036	-0.017	0.054	0.017	0.013	-0.063	1.000	0.271	0.224	0.262	0.205	0.166
Topic2	-0.027	-0.065	-0.037	0.106	0.027	0.031	-0.083	0.271	1.000	0.310	0.277	0.275	0.260
Topic3	-0.031	-0.059	0.018	0.048	0.012	0.036	-0.081	0.224	0.310	1.000	0.311	0.382	0.143
Topic4	-0.027	-0.055	-0.023	0.082	0.024	0.020	-0.066	0.262	0.277	0.311	1.000	0.280	0.206
Topic5	-0.024	-0.050	0.001	0.061	0.006	0.024	-0.083	0.205	0.275	0.382	0.280	1.000	0.123
Topic6	-0.026	-0.055	0.024	0.031	0.015	0.056	-0.086	0.166	0.260	0.143	0.206	0.123	1.000