UNIVERSITÉ LIBRE DE BRUXELLES
Faculté des Sciences
Département d'Informatique

Churn Prediction and Causal Analysis on Telecom Customer Data

Théo Verhelst

Promotor :
Prof. Gianluca Bontempi

Mémoire présenté en vue de
l'obtention du grade de
Master en Sciences Informatiques

Academic year 2018 - 2019

Abstract

Telecommunication companies are evolving in a highly competitive market where attracting new customers is more expensive than retaining existing ones. Retention campaigns can be used to reduce the incentives of churn, but this requires efficient churn prediction models. In this master thesis, we approach this problem with Orange Belgium customer data. A descriptive analysis of the dataset is conducted, and predictive modelling of churn is achieved with a random forest classifier and the Easy Ensemble algorithm. We assess the impact of different data preprocessing techniques such as feature selection and creation of new features. The directionality of the impact of variables on churn is estimated through a sensitivity analysis. Also, we explore the application of data-driven causal inference, which allows to infer causal relationships between variables purely from observational data.

Résumé

Les compagnies de télécommunication évoluent dans un marché hautement compétitif où attirer de nouveaux clients est plus coûteux que retenir les clients déjà présents. Des campagnes de rétention peuvent être utilisées pour réduire le taux de résiliation, mais cela nécessite des modèles de prédiction de résiliation efficaces. Dans ce mémoire de master, nous abordons le problème de prédiction de résiliation avec des données client de Orange Belgium. Une analyse descriptive du jeu de données est effectuée, et une modélisation prédictive de la résiliation est obtenue en utilisant un classificateur random forest et l'algorithme Easy Ensemble. Nous évaluons l'impact de différentes techniques de prétraitement de données, telles que la sélection de variable et la création de nouvelles variables. La directionalité de l'impact des variables sur la résiliation est déduite avec une analyse de sensibilité. Nous explorons également l'utilisation de l'inférence causale, qui permet de comprendre les liens de causalité entre différentes variables à partir de données d'observation.

Acknowledgements

...

Contents

Chapter 1

Introduction

1.1 Churn in the telecommunication industry

In recent years, the number of mobile phone users increased dramatically, reaching more than 3 billion users worldwide. The number of mobile phone service subscriptions is above the number of residents in several countries, including Belgium (**itu2018ict**). Telecommunication companies are evolving in a saturated market, where customers are exposed to competitive offers from many other companies. Hadden, Tiwari, Roy, and Ruta (2007) show that attracting new customers can be up to six times more expensive than retaining existing ones. This led companies to switch from a selling-oriented to a customer-oriented marketing approach. By building customer relationship based on trustworthiness and commitment, a telecommunication company can reduce the incentives for their client to churn, therefore increasing benefits through the subsequent customer lifetime value.

One of the various marketing processes used to improve customer relationship is to conduct retention campaigns. This traditionally consists in selecting clients at random and offering them some a promotion or advantage. Typical promotions include a reduced invoice, free calls, SMS or data volume. However, it is well possible that the customers thus reached might never have planned to churn in the first place. While this of course not a problem for the customer, it would be far more beneficial for the telecommunication company to be able to focus the retention campaigns on risky customers, in the hope of preventing attrition that would otherwise occur if no action is taken. The problem of detecting churn can be addressed with *data mining*, by collecting data about customers and using this information to infer typical patterns exhibited by risky clients. This data-driven approach is nowadays taken by most major telecommunication companies, and a part of the data mining literature is devoted to churn detection. We will describe this approach further in the next section.

1.2 Churn detection and prevention

The churn prevention process (depicted in figure 1.1) starts by collecting data about the customers and creating a *historical database*. This data summarizes the calls, messages, internet usage and other actions performed by the customers. It also includes

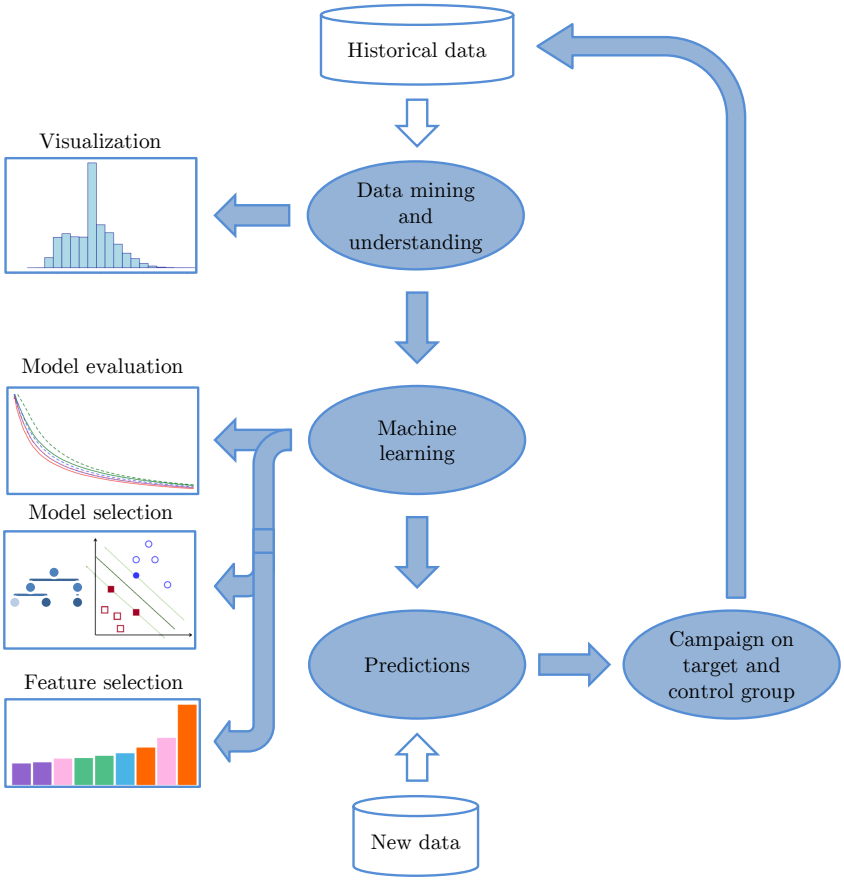


Figure 1.1 – The churn prevention process implemented at Orange.

information about the subscription, such as the type of tariff plan, its price, the subscription date, and so on. Finally, personal data provided by the customer may as well be used, such as the age or the place of residence. *Data mining* is then used to provide a quantitative understanding of the customers and their overall behavior. In particular, the difference between churners and non-churners can be highlighted. This is useful to decide which type of machine learning procedure should be used for churn prediction, but can also bring valuable knowledge to marketing and customer relationship teams. Once the data is sufficiently understood, a set of relevant *machine learning* models are built. These models learn from the historical data the patterns typically exhibited by clients that will churn in the near future. The types of patterns being learned, the techniques used to find them, and the underlying assumptions are dependent on the model under consideration. For example, a *naive Bayes classifier* assumes that each variable contributes to the probability of churn independently of the other variables. In order to decide which model should be selected, the performances of each model are compared by testing them on data left out of the training phase. Feature selection is also performed, and consists in evaluating the importance of each variable for predicting churn and training the model by using only the most important ones. This reduces the computation time, and generally improves the performances of the model since this reduces the noise unimportant variables bring along. Once an efficient model has been selected, the latest customer data is submitted to the model, which outputs a probabil-

ity of churn for each customer. By ordering the customers by churn probability, a list of the riskiest customers is established and sent to the campaigning team. They split this set of clients into a *target group* and a *control group*. Each customer of the target group is offered an incentive either by phone call, email or message, while the control group is left untouched. This allows, a few months later, to evaluate the impact of the retention campaign by assessing the difference of churn rate in retrospective in the two groups.

1.3 Causal inference

Depending on the resources available and the techniques used, this data mining pipeline can successfully predict potential churners, therefore allowing to conduct targeted retention campaigns. But campaigners are then faced with another challenge: what should they propose to the selected customers? Indeed, the predictions given by data mining algorithms usually just consist in a probability of churn. This prediction therefore does not indicate *why* the customer is about to stop her subscription. We need different analysis tools to tackle this problem. This is the purpose of *causal inference*, which is a formal approach to find the causes of some event in a system. Causal inference is usually conducted through *controlled randomized experiments*, but the scope of this master thesis did not allow such experiments. We therefore focus on data-driven approaches, which are based on some properties of the statistical distribution of causally linked variable.

To give an example of how causal inference can be conducted without experiment, consider the simplistic world where two tariff plans are available, an expensive one and a cheap one. The expensive plan makes customers consume more data per month, and increases their probability to churn. We represent such causal scenario with a *directed acyclic graph* as in figure 1.2. We plotted the data usage against the probability of churn for each tariff plan in figure 1.3. A positive correlation can be observed between data usage and churn, but disappears when considering each tariff plan separately. If a causal relationship also existed between data usage and churn probability, the correlation would still be visible, even when conditioning on the tariff plan. This idea of conditioning to discard putative causal links is at the basis of most causal inference algorithms.

Such inference methods are based on a number of assumptions, such as the absence of confounding factor. This assumption would be violated if, in our previous example, the age of the customer influences both its choice of tariff plan and its propensity to churn. The causal inference algorithm would still conclude that the expensive tariff plan causes clients to churn, while in reality they churn solely because of their young age. This sort of erroneous conclusions can lead to ineffective action in retention campaigns. Other inference methods imply other assumptions, and care must be exercised when using them.

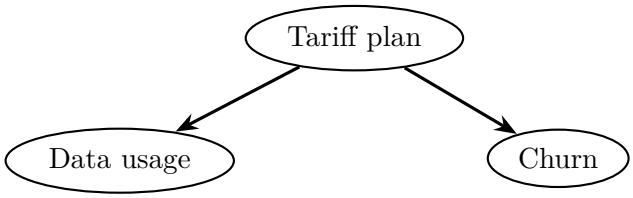


Figure 1.2 – Toy example of a causal diagram.

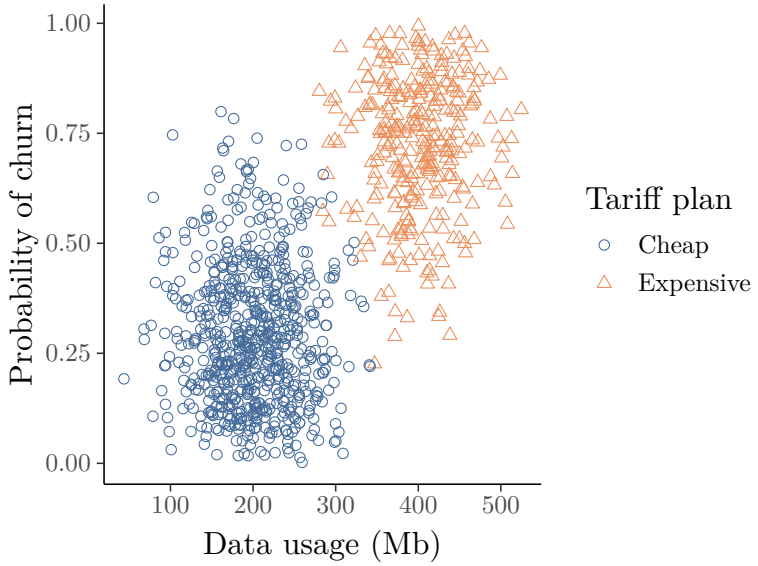


Figure 1.3 – Toy example of data usage against the probability of churn for different tariff plans.

1.4 Context and motivation

This master thesis is conducted in collaboration with Orange Belgium, and originated from the long-lasting scientific collaboration between Prof. Gianluca Bontempi (ULB Machine Learning Group) and Dr. Olivier Caelen (Orange Belgium). This collaboration enables us to work on real-world data, in Orange Belgium premises, and with people directly involved in the subject (data science and marketing teams at Orange). The churn prediction problem is challenging in many regards: the dataset is large (in our case, 1.1 million entries per month, on 6 months), highly imbalanced (there is very few churners in the whole customer base), and highly overlapping (many non-churners exhibit the same behavior as churners). It is therefore an interesting subject for a master thesis in machine learning, since it requires the use of different state-of-the-art tools for overcoming these challenges.

The interest for causal inference comes from the experience of the Machine Learning Group (MLG) in the subject. They mainly applied causal inference onto bioinformatics application, such as gene selection (Bontempi, Haibe-Kains, Desmedt, Sotiriou, & Quackenbush, 2011) or microarray data (Bontempi & Meyer, 2010). More recently, a competition in causal analysis was organized on the Kaggle website (https://www.kaggle.com/c/cause-effect-pairs), with the goal of fostering causal discovery between two variables. This led to the development of new methods, notably using

machine learning (Bontempi & Flauder, 2015). Moreover, the use of causal inference is seldom explored in the literature of churn prevention. It is thus stimulating to conduct research at the intersection of the these two domains, benefiting from the technical expertise of the MLG and the business knowledge of Orange.

1.5 Contributions

The main contributions of this thesis are

- Understanding of the churn prediction problem with a real-world dataset from a telecom company (section 3.1).
- Evaluation of the predictive power of a state-of-the-art churn prediction model, and several variations of the model using different features (section 3.3).
- Study of causal analysis, and its application to churn prediction from observational data (section 4.4).

1.6 Outline

This master thesis is partitioned into 5 chapters, presented here.

1. (Chapter 1) An introduction to the problem of churn in the telecommunication industry, the current methods in use to tackle it, and the contributions of this master thesis.
2. (Chapter 2) State of the art in churn prediction:
   - Choice of predictive model
   - Data preprocessing
   - Class balancing
   - Evaluation measure

   And state of the art in causal analysis:
   - Bayesian network learning
   - Markov blanket inference
   - Information-theoretic filters
   - Bivariate methods
   - Supervised methods
3. (Chapter 3) Assessment of a churn prediction model on Orange customer data. This chapter is further divided into:
   - Presentation of the dataset
   - Description of the data preparation

n	Number of features
$\mathcal{X} \subseteq \mathbb{R}^n$	Feature space
$\boldsymbol{X} = [X_1, \ldots, X_n]$	Vector of random variables in the feature space
$\boldsymbol{x} = [x_1, \ldots, x_n]$	Feature vector, realization of $\boldsymbol{X}$
$\mathcal{Y} = \{0, 1\}$	Target label space
$y \in \mathcal{Y}$	Example label
$P(e)$	Probability of an event e
$s \in [0, 1]$	Predicted churn probability for a given $\boldsymbol{x}$
$t \in [0, 1]$	decision threshold
$f_y(s)$	Probability density of s for instances labeled y
$F_y(s)$	Cumulated distribution of s for instances labeled y
$A \perp B$	Random variable A and B are independent
$\boldsymbol{X}_1 \setminus \boldsymbol{X}_2$	Set difference between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$

Table 1.1 – Summary of the mathematical notation.

- Description of the experimental setting
- Presentation of the results
- Comparison to other state of the art methods
- Conclusion and main outcomes of the experiments

4. (Chapter 4) Exploration of causal inference methods for churn prevention:
   - Theoretical background for causal inference
   - Scope of application
   - Description of the experimental setting
   - Presentation of the results
   - Discussion and conclusion
5. (Chapter 5) A conclusion, general remarks and directions for further work.

1.7 Notation

The mathematical notation used throughout this document is presented in table 1.1

Chapter 2

State of the art

2.1 Churn prediction

We organize our presentation of the state of the art in churn prediction by considering 4 aspects of interest in a usual churn prediction process: learning algorithm, data preprocessing, class balancing and choice of an evaluation measure.

Learning algorithm A large number of machine learning models have been applied to churn prediction in the literature. While some studies focus on simple and interpretable models (Dahiya & Bhatia, 2015; Keramati et al., 2014), other studies prefer the use of more complex models, at the expense of direct interpretation. These methods include boosting applied on simple classifiers (Vafeiadis, Diamantaras, Sarigiannidis, & Chatzisavvas, 2015), random forests, support vector machine, gradient boosting, among others (Umayaparvathi & Iyakutti, 2016; Verbeke, Dejaeger, Martens, Hur, & Baesens, 2012). An extensive overview of the current trends in churn prediction models is given in (Kayaalp, 2017).

Data preprocessing While the choice of model is important, it is equally important to consider a proper choice of features, data preprocessing and evaluation measure. Feature choice is limited by the available data infrastructure and usually consists of call detail records on the course of a few weeks. (Huang 2012) presented however how new kinds of features, including demographics profiles, marketing segments, and complaint information, can improve prediction accuracy.

Data preprocessing refers to the process following data acquisition and consists of modifying the data in various ways before the use in a predictive model. This comprises, non-exhaustively, feature engineering, data reduction, anomalies removal, encoding of categorical variables, and data normalization (Zhang, Zhang, & Yang, 2003). Coussement, Lessmann, and Verstraeten (2017) give an overview of common preprocessing steps used in churn prediction, and how careful preprocessing can positively affect the performance of the model. They even show that a simple logistic regression model applied to optimally preprocessed data competes with complex learning algorithms such as neural network or support vector machine applied on data preprocessed with a basic preparation procedure.

Recent years have witnessed the widespread usage of network-based classification for churn prediction. Verbeke, Martens, and Baesens (2014) present some experiment in this area of research. (Óskarsdóttir et al., 2017) perform an extensive benchmark of the different techniques proposed in the literature. The approach is based on call detail records (CDR) containing the communication logs of the subscribers. This data can be organized as a graph where nodes represent customers and edges represent social ties between customers. The basic assumption of network-based classifiers, which use such graphs to classify customers as churners or non-churners, is that customers having social ties with churners are more likely to churn themselves. This assumption is purposely loosely defined, as its exact implementation implies different assumptions and modeling decisions (Óskarsdóttir et al., 2017). From that, one can either construct a predictive model that directly uses the social graph, or extract features from the network and use them as the input of a conventional classifier, or even combine the two approaches. The outcome of the two articles is that relational and classical non-relational classifiers detect different types of churners and that a combination of both types of classifiers approaches performs best.

Óskarsdóttir, Van Calster, Baesens, Lemahieu, and Vanthienen (2018) present a novel, end-to-end approach to the problem of churn detection where a time-varying social network of the customer base is constructed, and a multivariate time series is then extracted for each customer from this network. Then, different time series classifiers are used to predict churn. A novel multivariate time series classifier is proposed, an adaptation of the similarity forest classifier (Sathe & Aggarwal, 2017). Óskarsdóttir et al. (2018) conclude that their approach outperforms state-of-the-art time series classifiers and non-time-based models for early churn prediction. However, static random forest and logistic regression are better at predicting late churn (that is, on short time scales).

Class balancing It is important to consider class imbalance when designing models for churn prediction. Indeed, the number of churners is usually very low compared to the total number of customers, and most machine learning models are usually not suited to handle highly imbalanced data (Batista, Prati, & Monard, 2004). Class balancing techniques have to be used to tackle this problem. These techniques can roughly be divided into two categories: data-level balancing and model-level balancing. Data-level balancing consists in modifying the dataset by either decreasing the number of majority instances, increasing the number of minority instances, or both. Model-level balancing consists in modifying the learning algorithm in such a way that minority instances are given more importance, often through the use of asymmetric misclassification costs (Zhu, Baesens, & vanden Broucke, 2017).

Evaluation measure The last step of a predictive model assessment is the evaluation of classification performance. Several performance measures exist for this purpose, such as precision, recall, F-score, area under the receiver operating characteristic curve (AUC), lift, maximum profit criterion, and expected maximum profit criterion. We briefly explain here these measures and their current use in the literature.

Precision is the fraction of true churners among all those we predicted to be churners, and recall is the fraction of predicted churners among all the true churners in the population.

$$precision = P(y = 1|s > t)$$
$$recall = P(s > t|y = 1)$$

In order to optimize both of these scores at the same time, one can use the F-score (also called F1 score, or F-measure), which is the harmonic mean of precision and recall.

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

By using the harmonic mean, we punish low values in any of the two values. It is sometimes used in the churn prediction literature (Ahmed & Maheswari, 2017; Keramati et al., 2014).

Receiver operating characteristic (ROC) curve (Krzanowski & Hand, 2009) is a plot of all possible compromises between true positive rate (TPR) and false positive rate (FPR). TPR is another name for recall, and FPR is the fraction of non-churners falsely predicted to be churners, among all non-churners.

$$FPR = P(s > t|y = 0)$$

In order to compare the ROC curve of different models, we use the area under the curve (AUC) as a measure of performance.

$$AUC = \int_{-\infty}^{+\infty} F_1(s) f_0(s) ds$$

AUC is very often used in churn prediction (Coussement et al., 2017; Mitrović, Baesens, Lemahieu, & De Weerdt, 2018) because it is less sensitive to class imbalance (few churners in a large population of customers) and misclassification cost (Verbeke et al., 2012).

An important drawback of these performance measures is that they do not represent the real objective of churn prediction: given that we are certainly not able to offer an incentive to all customers likely to churn, we have to restrict the retention campaign to a limited number of customers. From that, we wish to minimize the number of misclassifications occurring in this small subset of customers. This is the definition of the lift: given a certain set of customers (usually the subset of the customers with the highest predicted probability of churn), the lift is the ratio between the churn rate in this set and the churn rate in the whole customer base.

$$lift(t) = \frac{P(y = 1|s > t)}{P(y = 1)}$$

The value of the lift indicates how much we do better than choosing customers at random. For example, a lift of 3 indicates that the model will give a set of customers with 3 times as much churners as if we picked that many customers at random. The number of customers to consider is dependent on the application (it should ideally be

the number of customers reachable by the retention campaign), but a lift at 10% is sometimes used as a baseline (Verbeke et al., 2014; Zhu et al., 2017).

In order to determine formally how many customers should be included in the retention campaign, and therefore in the lift measure, Verbeke et al. (2012) developed the maximum profit criterion (MPC) and the expected maximum profit criterion (EMPC) (Verbeke et al., 2012; Verbraken, Verbeke, & Baesens, 2013). These two measures consist of evaluating the expected costs and benefits of conducting a retention campaign and selecting the optimal number of customers to call. The difference between MPC and EMPC is that MPC considers the cost and benefits to be known and fixed, while EMPC assigns a probability density function to these parameters. MPC and EMPC are often used in churn prediction (Óskarsdóttir et al., 2018; Stripling, vanden Broucke, Antonio, Baesens, & Snoeck, 2018; Zhu et al., 2017).

2.2 Causal analysis

Finding and using causes is crucial in human reasoning and decision making. While a predictive model returns the probability of a target value given that we observe a certain input vector, a causal model is supposed to return the target probability given that we set (e.g. by intervention) that input. The aim of causal analysis is to determine the consequences of manipulations and is opposed to the process of making predictions from observations. When used as a feature selection criterion, it enables increased robustness to violation of the iid assumption (e.g. concept drift) (Guyon, Aliferis, et al., 2007) and an enhanced understanding of the mechanism underlying the data. The gold standard for causal modeling is to carry out *randomized controlled experiments* (Fisher, 1937). For example, in order to assess the influence of moderate wine consumption on heart disease, it is not enough to measure wine consumption and heart disease in the population. This may lead to erroneous conclusions, such as a socio-economic factor that causes both increased wine consumption and risks of heart disease. In order to avoid such a problem, one may assign to a randomly chosen group of people a moderate wine consumption. If this group then shows a significantly different risk of heart disease, then we may conclude that a causal relationship indeed exists, and the apparent correlation is not due to a confounding factor. However, such experiments may be expensive, unethical, difficult to implement or unfeasible. This, along with advances in computation, data storage capabilities and new machine learning techniques, led to the development of causal inference based on observational data. We review here 5 approaches to data-driven causal inference:

- Bayesian network learning
- Markov blanket inference
- Information-theoretic filters
- Bivariate methods
- Supervised methods

All the approaches model causal relationships between random variables. These random variables represent, for example, the different features available about a customer in the churn prevention problem.

Bayesian network learning A causal Bayesian network is a discrete acyclic graph where the set of nodes correspond to the set of random variables, and a directed link indicates a causal relationship between two variables. This graph is also accompanied by the joint probability distribution of the set of random variables. Two conditions are usually imposed on the graph and the probability distribution (the causal Markov condition and the causal faithfulness condition) to ensure that it respects the semantics of a causal model. Notably, these conditions also allow predicting the effect of a manipulation (Spirtes, 2010). Causal Bayesian network can be learned from observational data, and two types of procedures have been developed for that purpose. The first one consists in a search in the graph structure space, and optimization a fitness function. This approach is detailed in (Heckerman, 1998). The second one is based on independence tests between pairs of variables, and iteratively construct and orient edges until a valid class of causal graphs is found. The PC and FCI algorithms use this approach (Shafer 1995).

Markov blanket inference The Markov blanket of a given target variable in a Bayesian network is a minimal set of variables that are shielding the target variable from the influence of other variables in the network. A formal definition is given in (Guyon, Aliferis, et al., 2007). This subset of variables contains all the information needed to predict the target variable (that is, any additional variable would be redundant). Moreover, if the causal Markov and faithfulness assumptions hold, then this Markov blanket is the set of direct causes (parents), direct effect (children), and also the direct causes of the direct effects (spouses). Inferring the Markov blanket of a variable, as opposed to inferring the complete Bayesian network, is beneficial when the number of variables is large, such as in microarray data. Several algorithms exist for Markov blanket inference, notably KS (Koller & Sahami, 1996), GS (Margaritis & Thrun, 2000) IAMB (Tsamardinos, Aliferis, & Statnikov, 2003), HITON (Aliferis, Tsamardinos, & Statnikov, 2003), and MMPB (Tsamardinos, Aliferis, & Statnikov, 2003). These algorithms start with an empty Markov blanket, and search for parents, children, and spouses of the target variable by using conditional independence tests. A number of heuristics are used to speed up the search, and most of these algorithms also include a second phase where false positives are removed from the result.

Information-theoretic filters Usual filter variable selection methods, based on a variable ranking, suffer from several drawbacks. For example, multiple variables may all provide the same information about the target (the extreme case occurring when a variable is actually a copy of another). In a univariate ranking approach, all of these variables would be included in the result, even though one is sufficient. Another typical problem occurs when two or more variables are not very predictive when taken individually, but on the contrary are predictive once used conjointly. These notions can be formalized using information theory, and information-theoretic filters are designed to

avoid the exposed pitfalls. Another advantage of such filters is that they do not make any assumption about the statistical distribution on the variables, thanks to the use of mutual information as a dependency measure. State-of-the-art information-theoretic filters include mRMR (Peng, Long, & Ding, 2005), DISR (Meyer, Schretter, & Bontempi, 2008), REL (D. A. Bell & Wang, 2000), CMIM (Fleuret, 2004) and FCBF (Yu & Liu, 2004). These algorithms vary on whether or not they take complementarity into account, if they avoid the estimation of multivariate density and if they return a ranking of the variable (as opposed to an unordered set of relevant variables). These filters can also be designed to explicitly favor variables having a direct causal influence on the target, thanks to dependence relationships uniquely exhibited by children, spouses and parents of the target. This is the case of the mIMR (Bontempi & Meyer, 2010) and the MIMO (Bontempi et al., 2011) filters.

Bivariate methods In the last decade, there is an increased interest in telling cause from effect from observational data on just two variables. This task is based on asymmetric properties of the joint distribution of the two variables, and because of the limited number of possible causal configurations, usual classification metrics can be used to assess the performance of inference methods. Moreover, the nature of the problem allows using classical machine learning algorithms for cause/effect classification. This approach gained interest through the organization of public a public challenge on Kaggle (https://www.kaggle.com/c/cause-effect-pairs). This competition resulted in several novel solutions for cause-effect detection, and a general advance in the state of the art. The winner of the challenge, the team *ProtoML*, extracts a large number of features from the variable distributions, and uses a large number of models for classification. The second-ranked participant, *jarfo* Fonollosa, 2016, uses conditional distributions and other information-theoretic quantities to infer features that are used for prediction. The general outcome of this challenge is that asymmetries in causal patterns enable the development of bivariate causal distinguishers, with an accuracy significantly better than random.

Supervised methods In the context of the aforementioned cause-effect pair competition, Bontempi and Flauder (2015) proposed an algorithm using asymmetries in the conditional distributions of the variables. They extended their method to a setting with more than two variables, by also extracting distribution features from other variables and using them to infer the existence and direction of a causal link in the two initial variables. This benefit is striking in the case of a collider configuration $x_1 \to x_2 \to x_3$: in this case, the dependency (or independence) between x_1 and x_3 tells us more about the link $x_1 \to x_2$ than the dependency between x_1 and x_2. By using a learning machine such as random forest, these features can be used to successfully infer causal links, competing with and often outperforming state-of-the-art Bayesian network inference methods.

Chapter 3

Churn prediction

3.1 Data

The data used throughout this work is a monthly summary of customers' activity, including mobile data usage, calls, and messages, along with information about the type of subscription, hardware, and socio-demographic information, for a total of 73 variables. The dataset comprises one entry per customer and per month, and a total of 5 months are present, from June 2018 to October 2018. About 1.5 million entries are present per month, for a total of about 7.6 million entries in the entire dataset. Also, in this dataset only private individuals having a non-zero usage of mobile data are present, this therefore does not represent the entire customer base of Orange Belgium. The target variable, churn, is represented as a date if the client is known to have churned, or an empty value otherwise.

Are present in this dataset two kinds of contracts, that we will call *SIM only* and *loyalty*. The first type refers to a subscription where the customer can churn freely at any time. This differs from the second type where the customer receives a large discount on the purchase of a mobile phone, but agrees not to churn for a certain period of time, usually 24 months. If the customer decides nonetheless to stop his subscription before the term of the contract, she has to pay back the remaining discount. After the data preprocessing step (presented in section 3.2, we are left with a total of about 5 millions entries corresponding to SIM only contracts, and about 250,000 entries corresponding to loyalty contracts. Therefore, we will mainly focus on SIM only contracts, for its broader impact on the customer base and its increased statistical significance. Some of the experiments will nevertheless be conducted on both types of contracts, in order to understand the differences in the churn dynamic.

Note that the number of entries in the loyalty dataset does not correspond to the real number of customers having this type of contract. This is due to the way the loyalty data entries are filtered. We focus only on customers approaching to the end of the mandatory period of the contract, since it is at this point that the churn rate raises. Before this period of time, the churn is almost non-existent, due to the remaining discount to be paid back by the customer. We consider a time frame of 2 months, explaining the low number of data entries under consideration.

The 73 variables are grouped into 6 categories:

- Subscription (17 variables)
- Calls and messages (11 variables)
- Mobile data usage (16 variables)
- Revenue (14 variables)
- Customer hardware (6 variables)
- Socio-demographic (5 variables)

The 4 remaining variables are the churn date, the timestamp of the data entry and two customer identifier columns.

Due to commercial reasons, a non-disclosure agreement prevents us from communicating precise details on the variables being used or the rate of churn in the customer base of Orange. The churn prediction problem is highly imbalanced, but we cannot disclose the exact ratio between churners and non-churners. The name of the different variables is limited to a letter corresponding to one of the 6 aforementioned categories and a number to differentiate variables among a category. We are able to disclose the exact name of a variable when its meaning is relevant for the discussion and its disclosure would not have a negative impact on confidentiality.

Most variables are continuous, such as the duration of phone calls, or the amount paid on the last bill. There are however a few discrete variables, either taking integer or categorical values. Such variables include the province of residence of the customer, or the number of active contracts. We present in this section a descriptive analysis of some variables, in order to understand, prior to any machine learning modeling, how they are distributed, and how they interact with each other. General patterns are qualitatively discussed in the text, and the figures support the discussion whenever possible.

One of the categorical variables represents the proportion of customers having a cable subscription besides their mobile phone subscription, such as for television or landline phone. There is less cable subscriptions among churners, and this fact is well known for Orange Belgium: a client having the cable will be less willing to churn, since this represents a significant investment in money and time. Another variable corresponds to the payment type for the bill. The two main types are automatic debit and bank transfer. We observe that less people among the churners chose an automatic debit. Although this is only speculation, this might be caused by the "bill shock" effect: when a client consumed more calls, messages or data than provisioned by its tariff plan, she has to pay an extra amount of money called out-of-bundle (OOB). When this amount is large, the client is more likely to get upset, and therefore is more likely to churn. But if her invoice is paid with an automatic debit, she may not notice this fact straight away, and this can therefore be a reducing factor of churn.

The bill shock effect is demonstrated in a more straightforward way on figure 3.1. The distribution of OOB is plotted for churners and non-churners on a logarithmic scale. There is a clear discrepancy between the two distributions, with churners having more often high amounts of OOB. This fact is often used to establish expert rules when

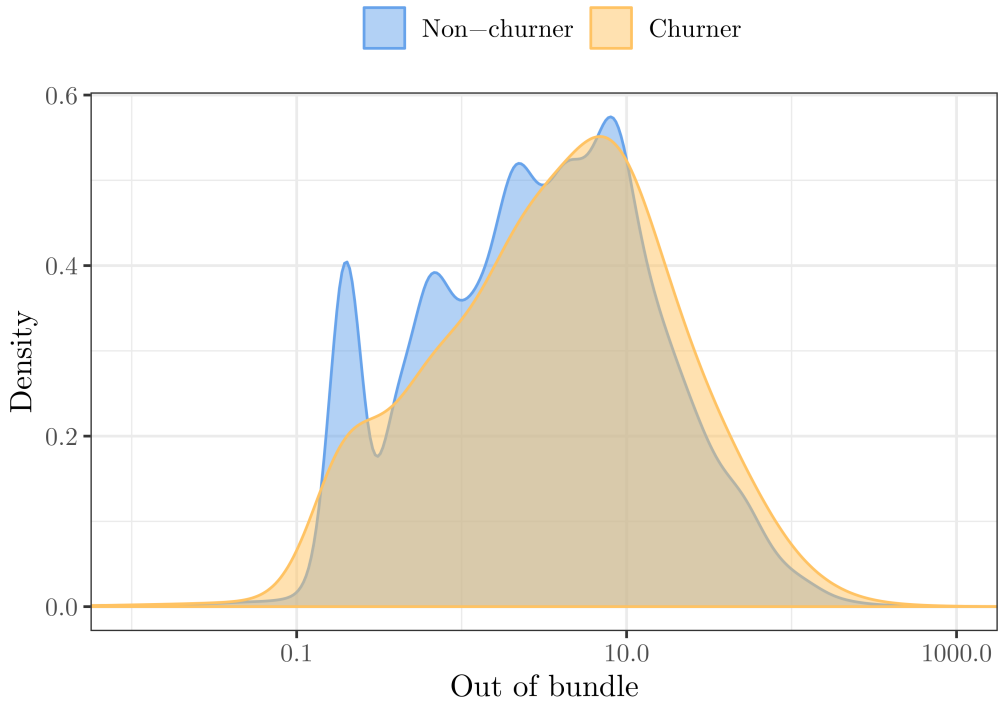


Figure 3.1 – Out-of-bundle amount (extra amount to pay on top of the usual invoice), in logarithmic scale.

conducting churn retention campaign: if a customer is likely to churn and has a large OOB, then a discount on the following invoices is offered, in order to mitigate the bill shock.

The importance of the tenure (the duration of the current subscription) is shown on figure 3.2. This graph displays the density of clients as a function of the tenure. For confidentiality reasons, the scale of the x-axis is hidden. The curve can be divided into two components: the new customers and the long-term customers. One can clearly observe that proportionally more churners are present in the first component than in the second. This indicates that long-term clients tend to churn less, whereas new clients are much more risky.

Figures 3.3 and 3.4 show the distribution of two discrete categorical variables.

We demonstrate the interaction between two categorical variables on figure 3.5. The horizontal axis indicates whether a customer has a cable connection, and the vertical axis denotes the payment responsible flag. This flag is set to false only when someone else pays the bill of the customer, such as a parent. Most customers of Orange Belgium do not have a cable connection, and are responsible of the payment, as indicated by the radius of the spots. The color of the spots indicates the churn rate, with a lighter color denoting a higher probability of churn. The area is proportional to the number of clients in each category. The impact of both binary variables appears clearly, with a significant difference of churn rate between the two extrema. Once again, the precise value of churn rate cannot be disclosed.

A principal component analysis (PCA) demonstrates the important overlap between churners and non-churners (figures 3.6 and 3.6). The blue correspond to non-churners, while the orange and yellow represent the churners respectively in the validation and

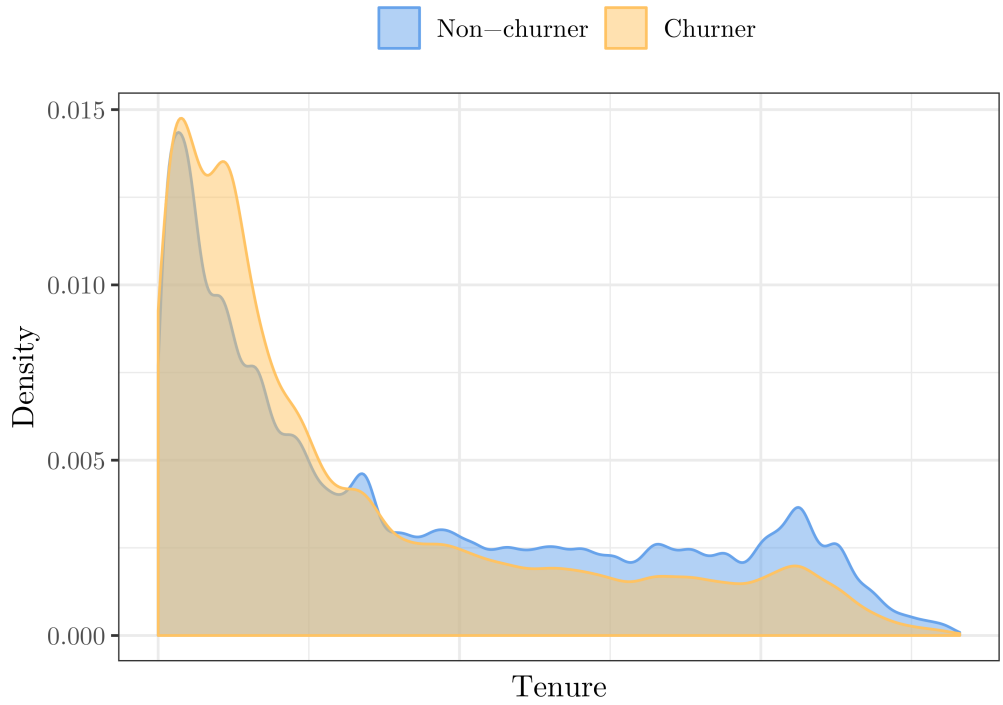


Figure 3.2 – Tenure (time spent without churning so far) for churners and non-churners.

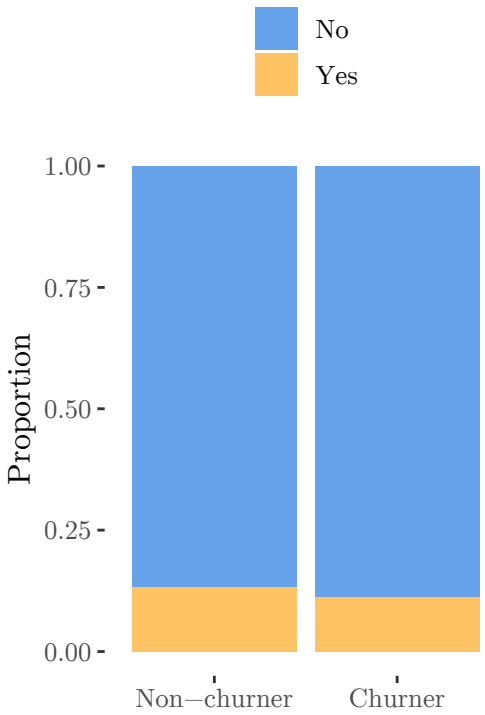


Figure 3.3 – Distribution of a binary variable related to hardware, for churners and non-churners.

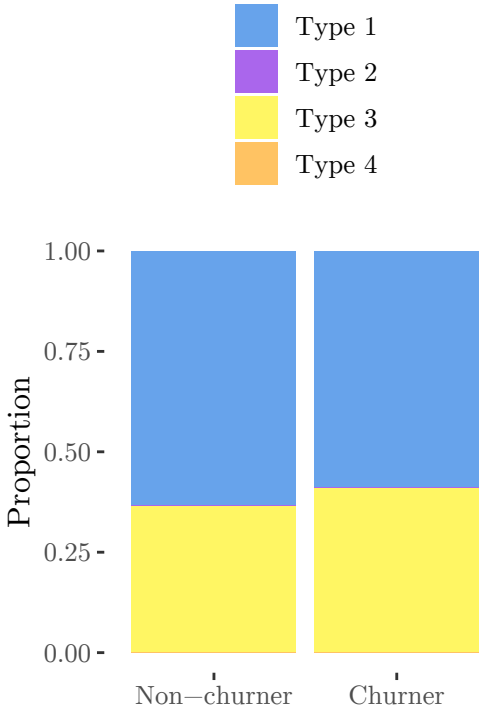


Figure 3.4 – Distribution of a categorical variable related to revenues, for churners and non-churners. 2 of the 4 possible values occur very rarely.

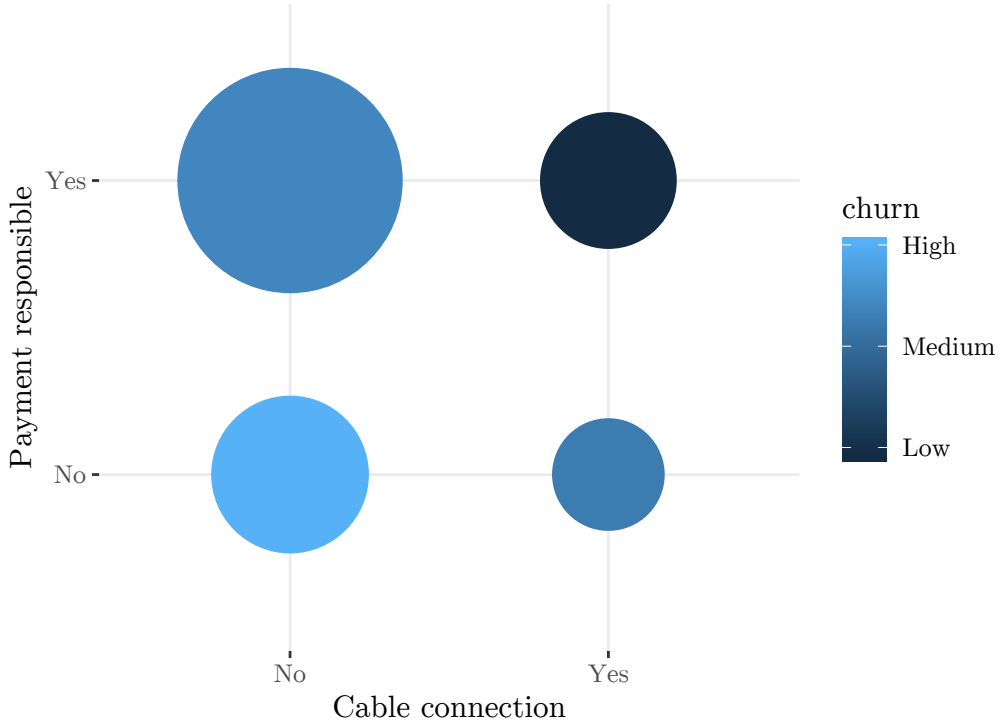


Figure 3.5 – Interaction between cable connection and payment responsible. A customer is not responsible of payment if someone else (e.g. a parent) pays the invoice in her stead. The color of the spots denotes the churn rate, whereas its area denotes the number of customers.

the test set. We explain how the test set and the validation set are partitioned in section 3.3. The ellipses represent the contour lines of covariance, that is, the set of points at a Mahalanobis distance of 1 from the mean in each set. The mean of each set is pictured as a dot in the center of the figure. It appears clearly that there is a large overlap between the population of churners and non-churners. Also, the standard deviation is larger in the population of churners, reflecting the interpretation that churn is associated with larger values for the out-of-bundle amount, number of calls, data usage, etc.

This data analysis demonstrates the high complexity of the churn prediction problem. No unique variable allows to unambiguously predict churn, since there is a significant overlap between the population of churners and non-churners. Informative variables such as the tenue (figure 3.2) or the out-of-bundle amount (figure 3.1) allow to increase or decrease the confidence into churn only marginally. A large number of variables must be used in conjunction in order to achieve decent predictive performances.

3.2 Data preparation

The data preparation is comprised of several steps conducted in sequence.

Unknown values preprocessing The original data uses various means to specify an unknown categorical value. For example, an unknown gender is either represented

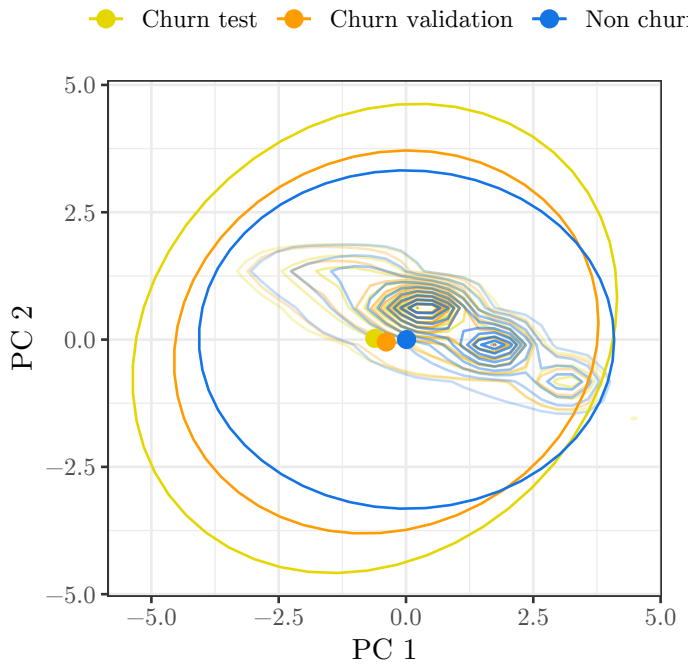


Figure 3.6 – Projection of the dataset onto the first two principal components. The ellipses show the set of points at a Mahalanobis distance of 1 from the mean of each group, which are represented by a dot.

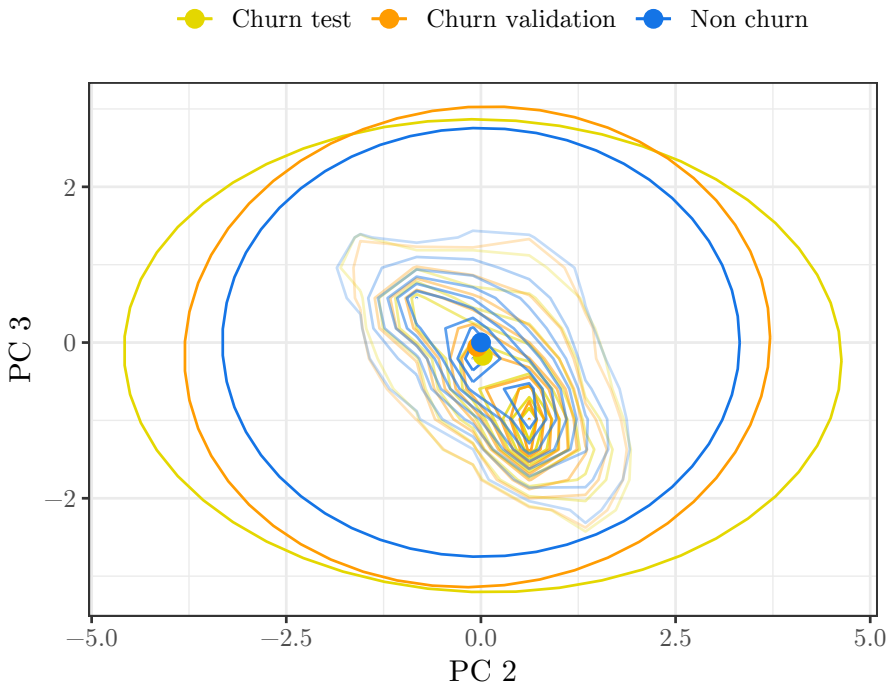


Figure 3.7 – Projection of the dataset onto the second and third principal components. The ellipses show the set of points at a Mahalanobis distance of 1 from the mean of each group, which are represented by a dot.

by an empty string, the string `"u"`, the string `"null"` or a null value. This preprocessing step replaces these various encodings by a unique one. Also, missing values in continuous variables are either replaced by zero or by the mean of the variable, depending on the semantics (e.g. null data usage is replaced by zero, whereas missing age is replaced by the average age in the dataset).

Date encoding Some fields are represented by a date, such as the last contract change, the date of contract activation, or the churn date. These fields are converted to the number of days between the first day of the month of the dataset entry and the field value. For example, let us consider an entry about the activity of a customer in January 2019. Say this entry contains a field with the contract activation date, with value `"20 December 2018"`. This field is converted to an integer value of 12, since there are 12 days between 1st January 2019 and 20 December 2018.

Clustering of character strings There are three character string variables, representing the current and the previous tariff plan, and the manufacturer of the customer's device. These three variables could be considered as categorical variables, but the high number of different values would make this difficult to implement. We alleviate this difficulty by clustering all the different values into a small number of groups. In the case of the two tariff plan variables, this corresponds to the different tariff plan options (*Hummingbird*, *Koala*, *Eagle*, etc). For the device manufacturer, we keep the 7 most common values, and we replace all the less frequent values by "Other".

Difference and ratio columns For each numerical field representing a quantity that can change from month to month (such as the total duration of calls, or the mobile data usage), we create 2 additional fields. They contain the difference and the ratio of the value of the field with that of the same field the previous month. This hopefully gives the model an indication of the customer's behavior evolution over the course of the last month. 41 variables are suitable for this operation, therefore increasing the number of variables up to 155. If no data is available for the previous month (such as for the first month of data), the differences are set to 0 and the ratios are set to 1. In order to reduce computation time, not all experiments use these new columns, as discussed in the next section. This augmented dataset is named "SIM only Δ" thereafter.

Normalization The data is normalized to obtain zero mean and unit variance. Even though the only models being used are random forests, which are not sensitive to linear scaling of its input variables, this step is kept in the event that we would have tried another model requiring normalization (such as support vector machine or neural networks). This preprocessing step is also useful for sensitivity analysis, where we add a small value to a variable and observe the difference in the predictions of the model. In this case, a normalized dataset allows to use the same difference value for all variables.

Target variable The last step is to create a binary target variable from the churn date. It is defined to be true if and only if the date of churn is in the two months

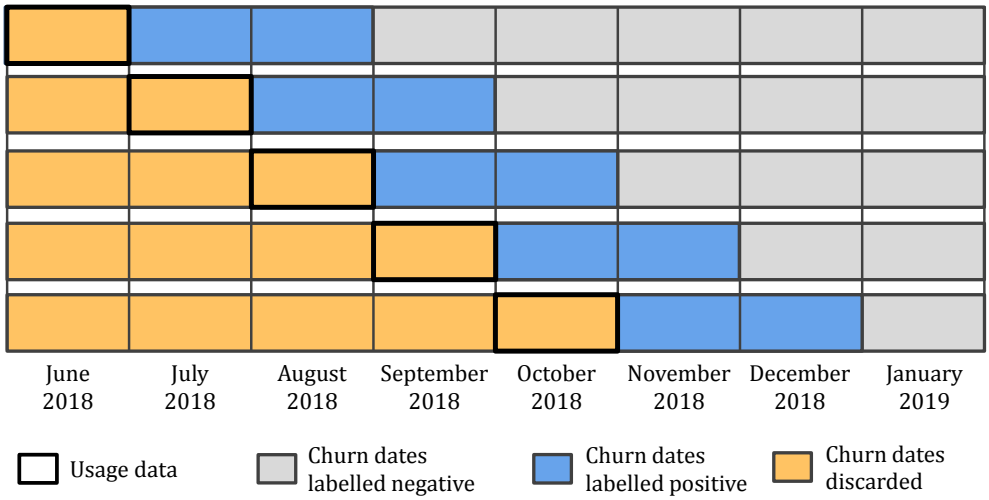


Figure 3.8 – Outline of the target variable assignment. The dataset is separated in 5 months, and each data entry in each month is labeled as churner if the churn date is given and less than two months ahead.

following the current data entry. If the churn date is in the current month or before, then the entry is discarded, for two reasons. Firstly, the information contained in these entries is incomplete. Secondly, this data is not relevant for churn prediction, as we wish to predict churn at least a few days in advance. It is not interesting to learn patterns exhibited by clients that will churn the next day, as there is probably no longer any hope of successful retention. If the churn date is not given, or if it is more than two months after the month of the data entry, the churn variable is set to false. This process is pictured in figure 3.8. The choice of the time threshold is dependent on the business application. A lower threshold focuses on short-term churn, whereas a larger threshold enables to predict churn from further in the future. Models already in use at Orange Belgium consider a time period of two months, we therefore use this value in order to enable the comparison of our results with production models.

3.3 Experiments

Scope

The experiments on predictive modeling consists in the training of predictive models on the data described in the previous sections, and an assessment of their performance. Three datasets are derived from the output of the preprocessing step: one containing the loyalty contracts, one containing the SIM only contracts, and one containing the SIM only contracts with difference and ratio variables (called SIM only Δ). We evaluate the impact of

1. variable selection, based on the feature importance provided by trained random forest models;

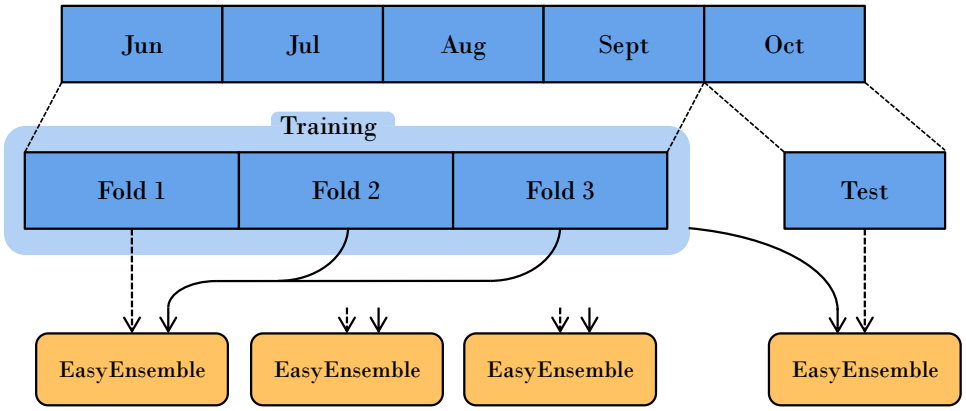


Figure 3.9 – Outline of the data repartition between training and test set, with a 3-fold cross-validation on the training set. Dotted arrows indicate testing and solid arrows indicate training. Arrows for two of the three validation models are not shown.

2. the addition of difference and ratio variables;

3. the type of contract (SIM only vs. loyalty).

The high computational cost of the model training on such a large dataset does not allow to test all the possible configurations of these three parameters. We limited the number of selected variables to 20, 30 or all variables. Also, we do not explore the difference variables for the loyalty contracts. This combinations of parameters yields 9 different experiment configurations.

Data segmentation

In each configuration, the corresponding dataset is split into a training and a test set. The training set comprises the first 4 months of data, from June 2018 to September 2018, and the test set comprises only the last month, October 2018, as pictured in figure 3.9. Separating by months allows for a potential concept drift from the training set to the test set, thus assessing better the generalization abilities of the model. This also indicates whether training has to be repeated each month as new data arrives from the customers. We perform a k-fold cross-validation on the training set in order to provide an indication of the performance of our model on the training data. We set $k = 3$, as a compromise between statistical significance and computation time. If the results obtained on the training set by cross-validation are significantly better than those on the test set, this indicates that the model generalizes poorly on unseen data. Note that when testing the model on the test set, a new model is trained on the whole training set.

Class balancing

In order to counteract the sheer prominence of non-churners in the dataset, we need to use class balancing. Given the set of positive instances $\mathcal{P}$ and the set of negative instances $\mathcal{N}$, usual undersampling methods consist in choosing a subset $\mathcal{N}' \subset \mathcal{N}$ such

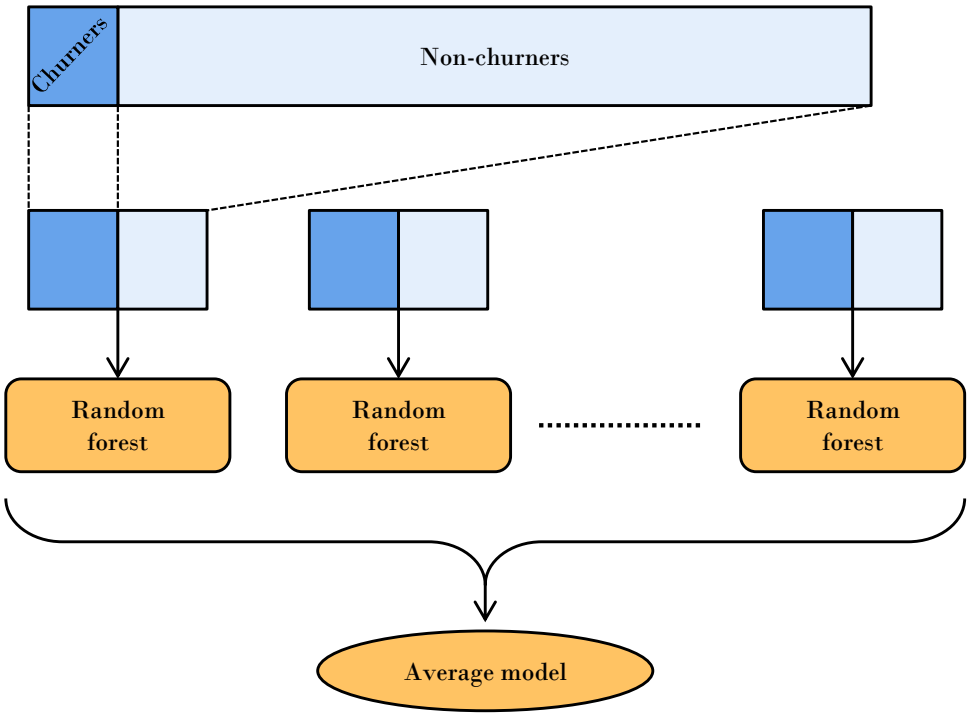


Figure 3.10 – Easy ensemble methodology for unbalanced data. A set of random forests is trained each on the whole set of positive instances, and on a randomly chosen subset of negative instances. The final predictions are the average of all the random forests individual predictions.

that, by training on $|\mathcal{N}'| \cup |\mathcal{P}|$, the ratio between positive and negative instances is suitable for a given learning algorithm. It is common to chose a ratio of 1. The main issue with this approach is that a large number of positive instances (i.e. all of $\mathcal{N} \backslash \mathcal{N}'$) is ignored. EasyEnsemble (Liu, Wu, & Zhou, 2009) overcomes this issue by independently sampling $T > 1$ subsets $\mathcal{N}_1, \ldots, \mathcal{N}_T$ from $\mathcal{N}$. Then, T predictive models are trained on all of $\mathcal{P}$ and individually on each $\mathcal{N}_i$. The final predictions are the average of the predictions of the T models. This process is pictured in figure 3.10. In our case, we use $T = 10$, and the predictive models are random forests.

Evaluation measure

The performance of the different models are evaluated using three different measures: the lift curve, the receiver operating characteristic (ROC) curve, and the precision-recall (PR) curve. While the ROC curve and the PR curve are widely used in the machine learning literature, the lift curve is of more practical interest for evaluating churn prediction. Since a customer churn retention campaign focuses on a limited amount of customers, the lift curve allows to observe the expected performance of the model as the number of customers included in the campaign varies. From the ROC and PR curves, we derive the area under the ROC curve (AUROC), the area under the PR curve (AUPRC) and the lift at different thresholds (1%, 5%, and 10%).

We argue that the most sensible cost evaluation functions are the maximum profit

criterion (MPC) and the expected maximum profit criterion (EMPC) (Verbeke et al., 2012; Verbraken et al., 2013). These take into account the different costs and benefits yielded by a retention campaign, and provide the decision threshold that should be applied to maximize the profit given the probability distribution output by a prediction algorithm. This process formalizes the intuition behind the lift criterion that the prediction algorithm should focus on reducing false positives, since the retention campaign is not able to reach each and every potential churner. Despite the relevance of this approach under a profit-centric point of view, we are not able to use this evaluation measure in our study. This is caused by the necessity to evaluate different costs and benefits parameters, such as the cost of reaching a customer, the probability that a customer accepts an incentive, the benefit if the customer accepts it, etc. The evaluation of these parameters is a time-consuming process, and is outside of the scope of our work.

Sensitivity analysis

The impact of variables on churn prediction is derived in two different ways. The first corresponds to the variable importance output by the random forest models. Each random forest calculates a score for each variable by measuring how much the prediction accuracy decreases when all the values of this variables are randomly permuted. The decrease in accuracy is calculated over the out-of-bag samples in each tree. This permutation cancels out any statistical dependency between this variable and the target variable, giving an estimate of the importance of the variable in the trained model. Note that if two variables share the same information about the target (for example by being highly correlated), the importance of both of these variables will be less that if only one were present. This is due to the fact that the two variables are equally likely to be chosen when splitting nodes in a tree, therefore reducing the impact of the removing one of the two variables when computing the importance.

This measure of importance allows to understand the predictive power of each variable, but does not indicate the directionality of its impact on the predictions. We address this issue by constructing, for each variable X_i, an alternate training set identical to the original one, but where a value equal to one standard deviation σ_{X_i} is added to each instance of the variable X_i. A second shifted dataset is also constructed by subtracting instead of adding the standard deviation. Then, the average predicted probability of churn is computed for both the original training set and the shifted one, and the difference between the two average probabilities is taken. This difference indicates the impact of the variable on the predictions. For example, a predicted churn probability lower for the training set where a standard deviation is added to the tenure variable indicates that longer-standing customers are associated with less churn. Note that we use in this experiment the normalized dataset, so that adding a standard deviation amounts to adding 1 to the instances of the variable.

	SIM only			SIM only Δ			Loyalty		
	20	**30**	**All**	**20**	**30**	**All**	**20**	**30**	**All**
AUROC	0.66	0.73	0.73	0.72	0.73	0.69	0.74	0.76	0.76
AUPRC	0.05	0.10	0.10	0.10	0.10	0.08	0.15	0.19	0.18
Lift at 10%	2.25	3.34	3.41	3.27	3.42	3.03	2.96	3.40	3.30
Lift at 5%	2.64	4.49	4.68	4.48	4.67	4.09	3.51	4.22	4.02
Lift at 1%	4.29	9.20	9.53	10.09	9.95	7.67	4.66	6.65	6.16

Table 3.1 – Summary of the results of prediction experiments on the test set. Highest values for each type of contract and for each evaluation measure are underlined for the test set.

3.4 Results

This section shows the results of the predictive experiments. Figures 3.11 to 3.19 are performance curves for the three different datasets. Each plot contains a curve for both validation and test sets, and for the configurations where we select 20, 30 or all of the variables. This amounts to a total of 6 curves per plot. As explained in section 3.3, the validation is done on the 4 first months of data, from June 2018 to September 2018, whereas the test set correspond to data from October 2018. Figures 3.11 to 3.13 show the lift curves, figures 3.14 to 3.16 show the ROC curves, and figures 3.17 to 3.19 show the precision-recall curves. A summary of the results is given in table 3.1. The lift at different thresholds, the area under the ROC curve, and the area under the precision-recall curves are reported for the test set. Table 3.2 provides the same information for the predictions on the validation set. The impact of the different experiment parameters on predictive performance is discussed in the next sections.

In this section, numerical scores associated to variables are displayed as horizontal bar plots. The colors of the bars correspond to the variables categories presented in section 3.1:

- Subscription
- Calls and messages
- Mobile data usage
- Revenue
- Customer hardware
- Socio-demographic

Number of variables

The number of variables has an influence on the prediction accuracy, and this impact depends on the dataset under consideration. Recall that in each configuration, the selected variables are chosen according to the variable importance given by the random

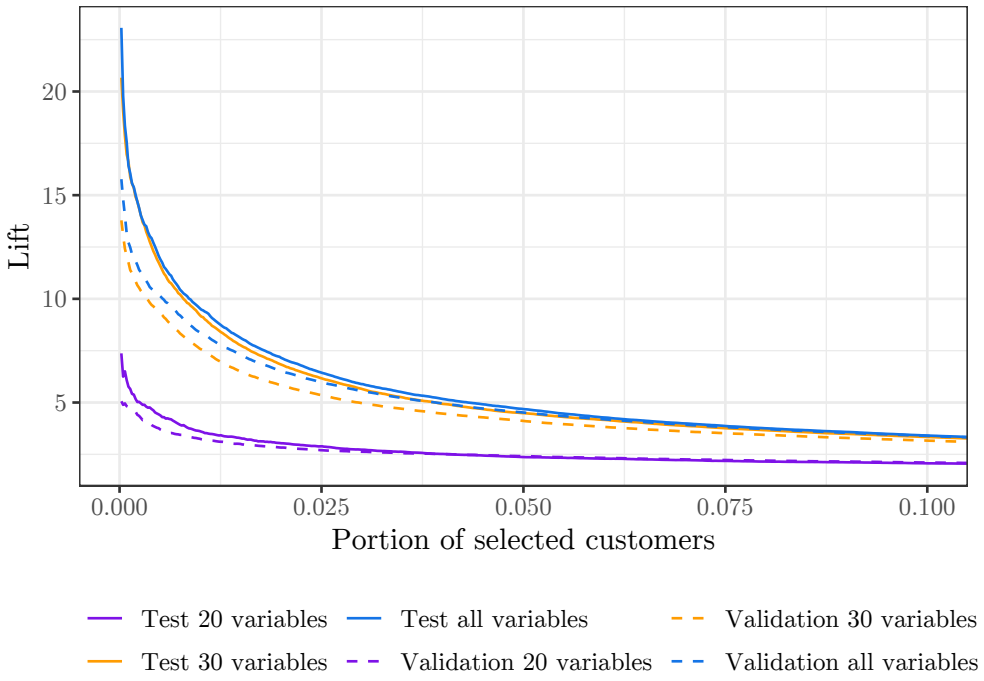


Figure 3.11 – Lift curve for SIM only

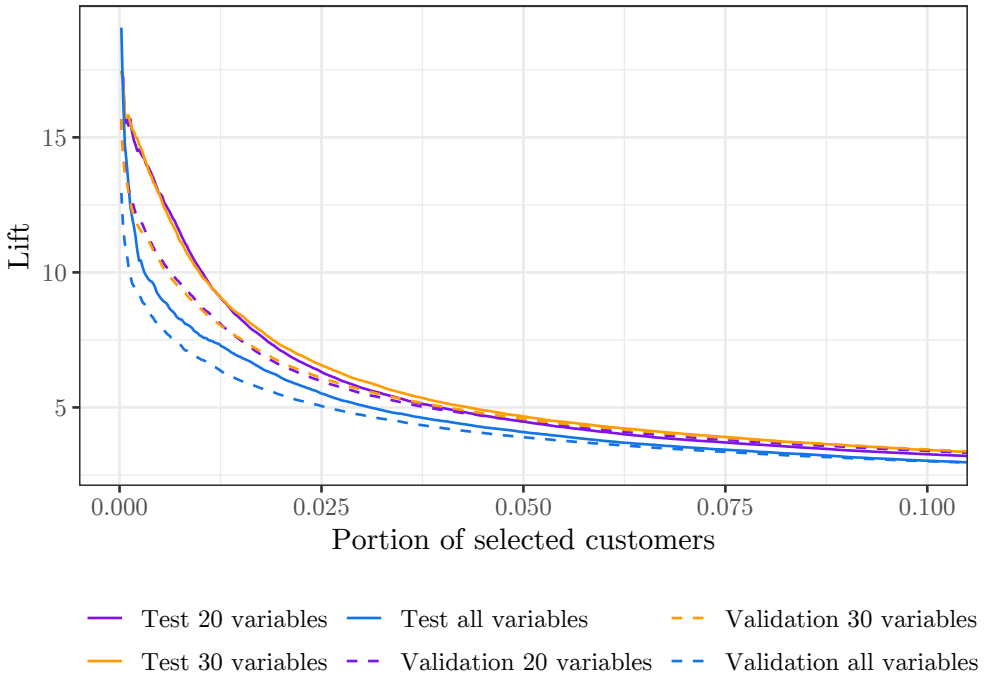


Figure 3.12 – Lift curve for SIM only with difference and ratio variables

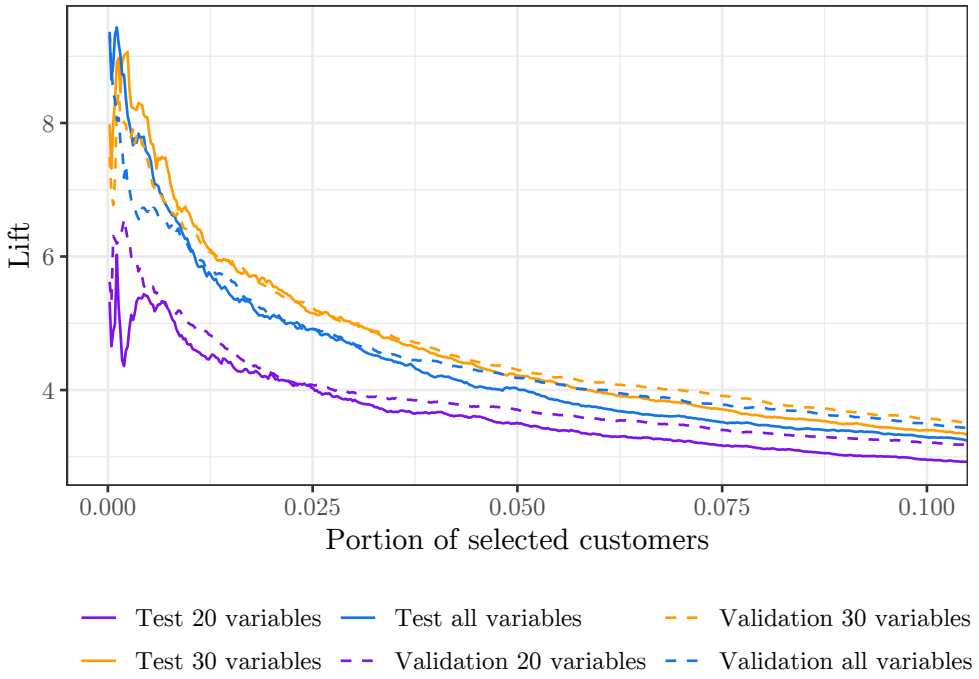


Figure 3.13 – Lift curve for loyalty

forests trained on the whole training set. In the case of the SIM only dataset (figures 3.11, 3.14 and 3.17), selecting only 20 variables decreases drastically the performance. Selecting 30 variables achieves performances almost as good as selecting the whole set of 73 variables.

For the SIM only Δ dataset, a lower number of variable is beneficial for performance. As shown on figures 3.12, 3.15 and 3.18, selecting all variables appears to be detrimental to the accuracy, whereas choosing between 20 or 30 variables does not make a significant difference. This might be caused by the additional variables that add more noise than useful information for the random forests. It is interesting to note that selecting the top 20 variables when considering the SIM only Δ dataset provides much better accuracy than selecting the top 20 variables in the original dataset. As shown in figure 3.21, the 20 most important variables in the SIM only Δ dataset do not even include any difference or ratio variable. The difference between the top 20 most important variables is as follow:

- SIM only Δ includes the number of contracts, the age, and a variable on data usage, whereas SIM only does not;
- SIM only includes the device manufacturer, the previous tariff plan, and two other variables on data usage, whereas SIM only Δ does not.

The large difference in accuracy between these two configurations must be caused by this difference in the selected variables, since all other variables and all other experiment parameters are identical.

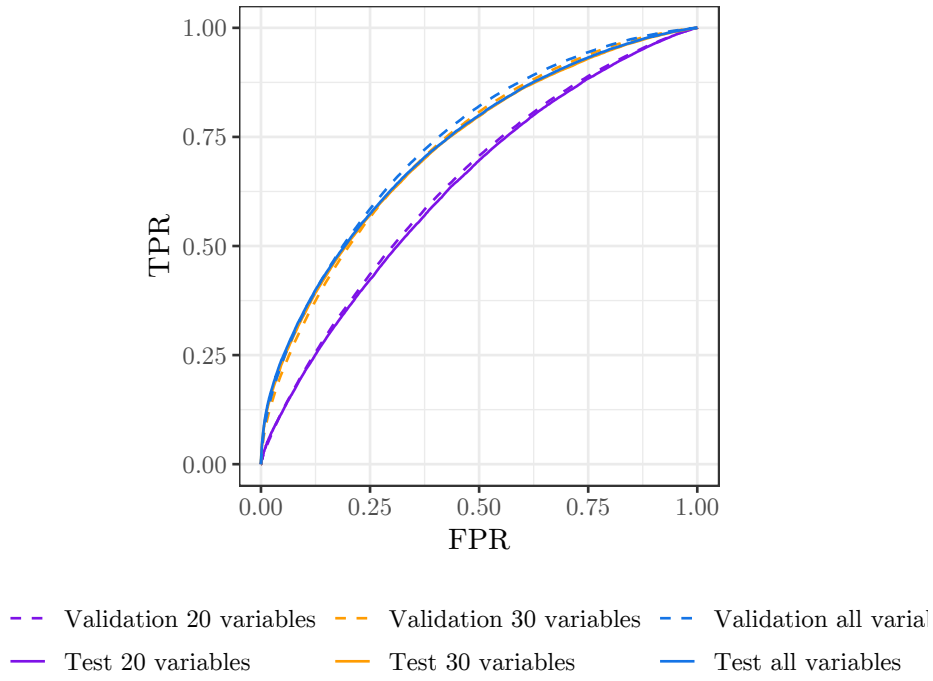


Figure 3.14 – ROC curve for SIM only.

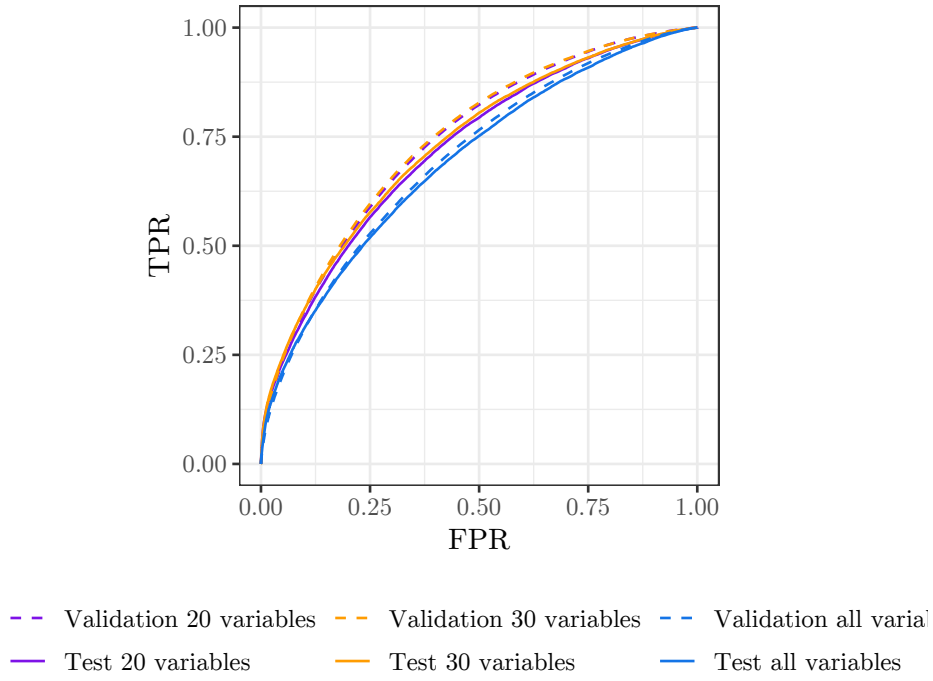


Figure 3.15 – ROC curve for SIM only with difference and ratio variables.

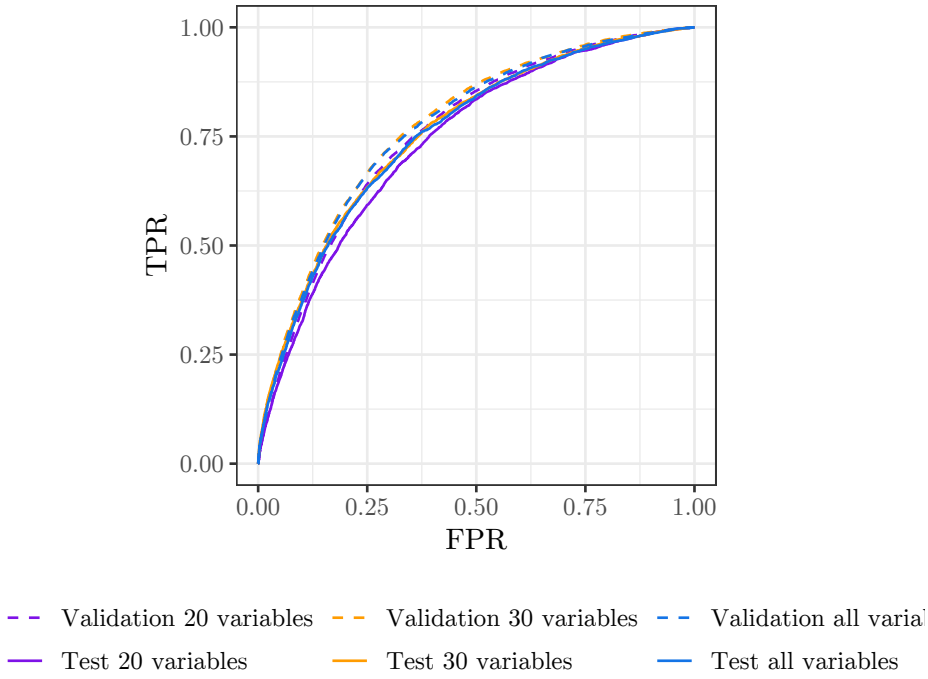


Figure 3.16 – ROC curve for loyalty.

When considering the loyalty dataset, selecting 30 variables instead of the whole set of 73 variables is marginally beneficial on low threshold (less than 0.1). On the overall space of thresholds, the difference is not significant, as indicated on table 3.1 with the AUROC and the AUPRC. However, selecting only 20 variables is clearly detrimental.

Difference and ratio variables

The difference and ratio variables do not have a positive impact on performance. Indeed, the best performance achieved on the SIM only Δ dataset are obtained by limiting the number of variables to 20 or 30. The variables being selected in these cases do not include any of the difference and ratio variables. If all variables are used, the performance of the random forests decreases significantly, as shown in figures 3.12, 3.15, 3.18, and table 3.1. Moreover, the memory usage of this dataset is much higher than that of the original dataset, thus complicating the training process.

Generalization performances

The generalization abilities of the trained models is evaluated by comparing the accuracy on the validation set and on the test set. On the lift curves and on the precision-recall curves, it appears that for all configurations, the performance on the test set is better than on the validation set. Bear in mind that on these curves, only a small fraction of the threshold space is represented, whereas the ROC curves show all possible decision thresholds. This observation implies that customers with a very high probability of churn are proportionally more numerous in the test set than in the validation set. It is illustrated on figures 3.6 and 3.7, where the ellipse of covariance of the test set

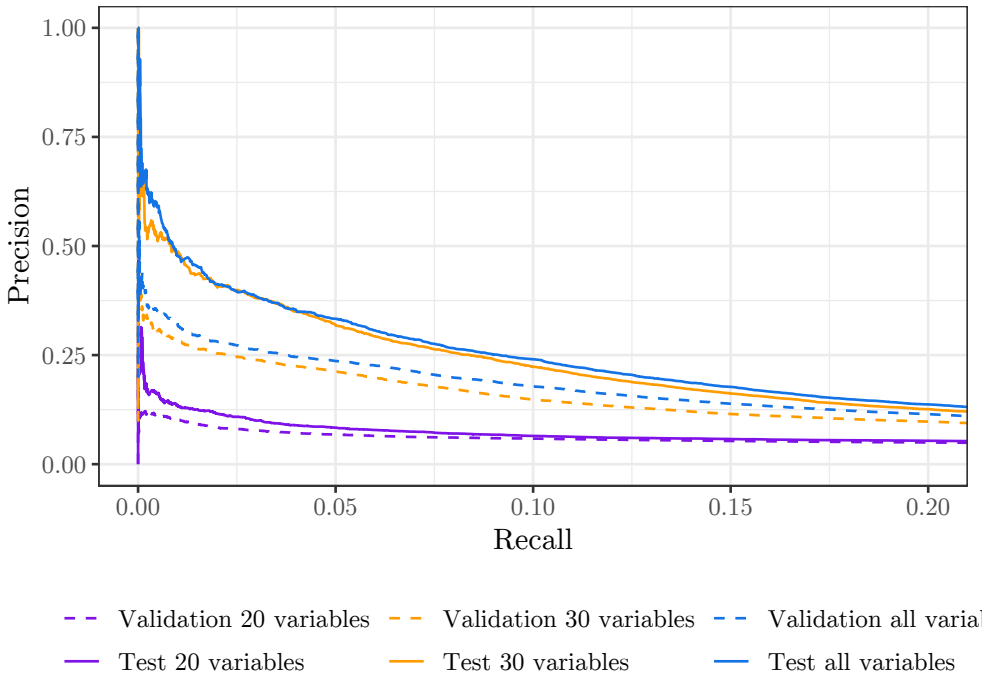


Figure 3.17 – Precision-recall curve for SIM only.

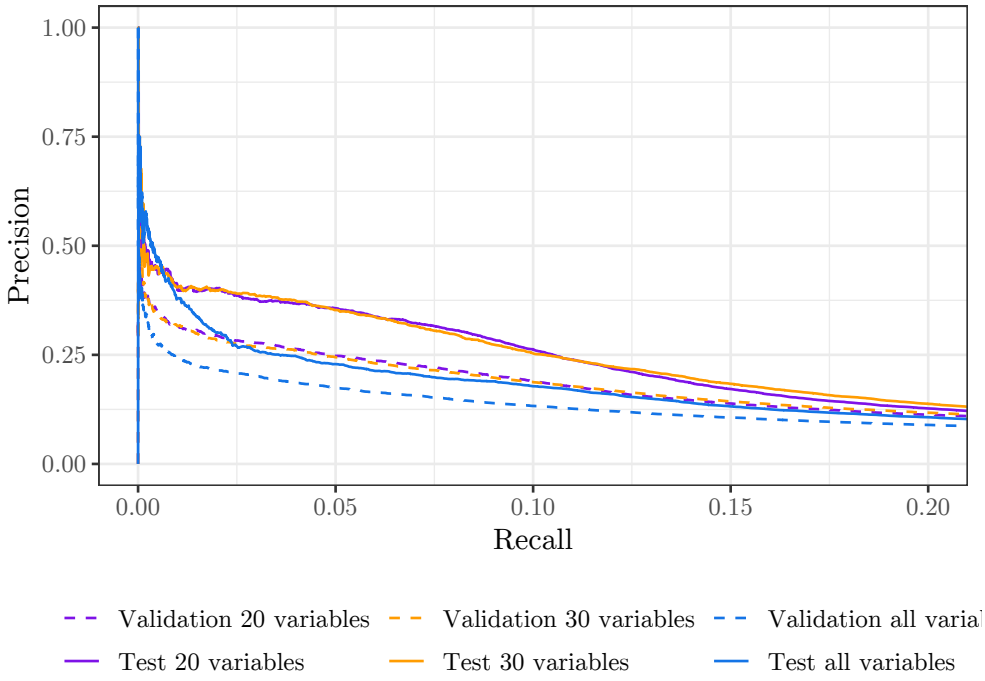


Figure 3.18 – Precision-recall curve for SIM only with difference and ratio variables.

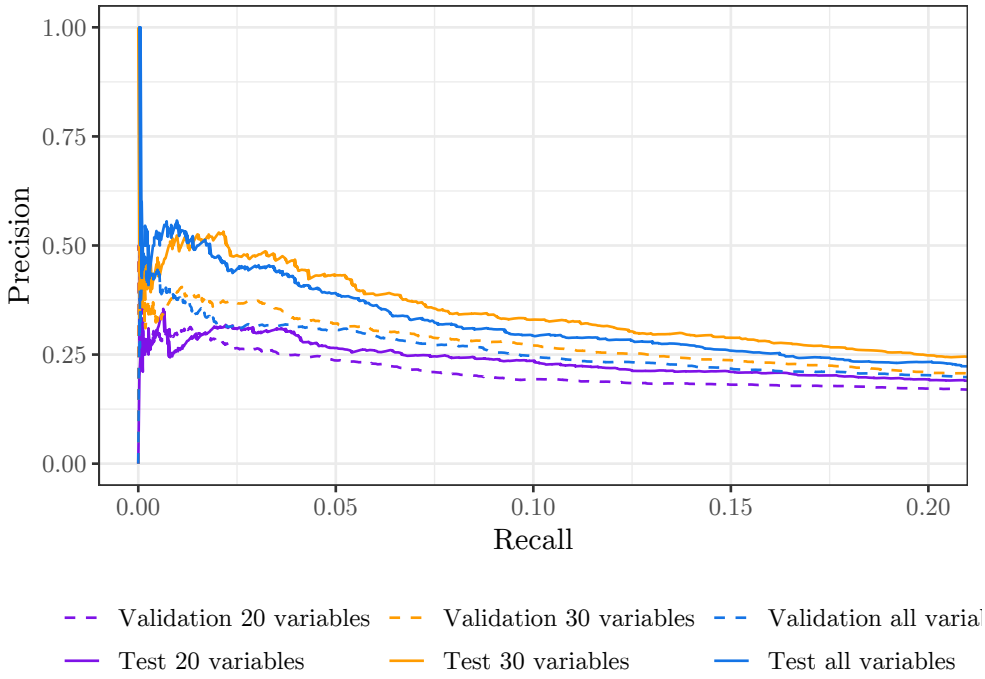


Figure 3.19 – Precision-recall curve for loyalty.

	SIM only			**SIM only Δ**			**Loyalty**		
	20	**30**	**All**	**20**	**30**	**All**	**20**	**30**	**All**
AUROC	0.64	0.73	0.74	0.74	0.74	0.70	0.76	0.78	0.77
AUPRC	0.04	0.08	0.08	0.09	0.09	0.07	0.13	0.16	0.15
Lift at 10%	2.10	3.16	3.39	3.39	3.44	3.01	3.22	3.57	3.50
Lift at 5%	2.41	4.11	4.52	4.49	4.57	3.90	3.71	4.30	4.18
Lift at 1%	3.24	7.58	8.36	8.80	8.67	6.79	5.00	6.37	6.11

Table 3.2 – Summary of the results of prediction experiments on the validation set.

is larger than that of the validation set. This suggests that, in the test set, there are more churners with very high values for variables having a large standard deviation. This probably corresponds to the bill shock effect discussed in section 3.1: a large out of bundle amount, caused by a large data consumption, increases the probability of churn. In this case, the bill shock is more pronounced in the test set, explaining the improvement in predictions.

The performance measures on the validation set are summarized in table 3.2. When we compare the lift at low thresholds in table 3.1, it is clear that the model manifests worse performance on the validation set than on the test set. However, it is not the case for the AUROC and the AUPRC, which take into account the whole space of decision threshold, and not only the most risky customers. We can conclude that our model has been trained on a training set where the overlap between churners and non-churners is more important than on the test set. The model thus generalizes well, and even perform better on unseen data in our case due to a lucky domain shift.

Type of contract

As indicated on table 3.1, the models trained on loyalty perform slightly worse than that of the SIM only datasets on small threshold, but better on larger thresholds. The AUROC is equal to 0.76 for loyalty, whereas the best performing configuration for SIM only achieve an AUROC of 0.73. The AUPRC is almost the double, and this can be seen on figure 3.19. The precision is similar to that of the SIM only for low recall, but decreases much more slowly. It is still at approximately 0.25 when the recall is 0.2, whereas in the SIM only PR curves, the precision is already at 0.12 at this threshold.

Recall that there is less loyalty customers than SIM only, less confidence can thus be given to the statistical results for loyalty, especially at low thresholds. According to these results, the models trained on the loyalty customers are slightly less efficient on small threshold, but this is made up for on larger thresholds. The increase in performance is probably due to the more obvious churn patterns exhibited by loyalty customers. Indeed, most of the churn in this population is due to the end of the mandatory period of the subscription. This is reflected on figure 3.22, where time-related variables are prominent in variable importance. Given that we know when the mandatory part of the customer's contract ends, the confidence of the model in the probability of churn is increased compared to the SIM only case.

Sensitivity analysis

The results of sensitivity analysis are shown in figures 3.20 to 3.24. Figures 3.20 to 3.22 show the variable importance given by the random forest models. There is one plot per dataset, and in each plot the importance of each variable is averaged over the 10 models underlying the Easy Ensemble meta-model. As discussed in the previous section, the most important variables for the SIM only and the SIM only Δ datasets are almost identical. They consist in a mix of socio-demographic variables (e.g. the province), information about the tenure and the tariff plan of the customer, and aggregate variables related to phone calls. The difference and ratio variables are ranked fairly low for SIM only Δ, the first one having a rank of 40 (not shown on figure 3.21). On the other hand, the selected variables for the loyalty dataset (figure 3.22) are quite different. The tenure (first and third variables) and other time-related variables on the subscription are all important variables. Information relative to the type and time of subscription is therefore important for predicting churn in loyalty contracts. This is illustrated by the yellow color dominating the graph. Also, the age is more important than for the SIM only dataset, as well as the variables U1, U2 and U3, corresponding to data usage. This is consistent with an interpretation of a younger and fickler customer base, consuming more data and more prone to churn.

Figures 3.23 and 3.24 display the shift in predicted churn probability when a variable is offset by one standard deviation. Figure 3.23 correspond to an increase in the value of each variable, whereas figure 3.24 correspond to a decrease. The tenure and the number of contracts have a symmetric effect: an observed increase in their value reduces the probability of churn, and conversely. For the tenure, this correspond to the intuition that a newer customer is more prone to churn. The number of contracts is a marker of

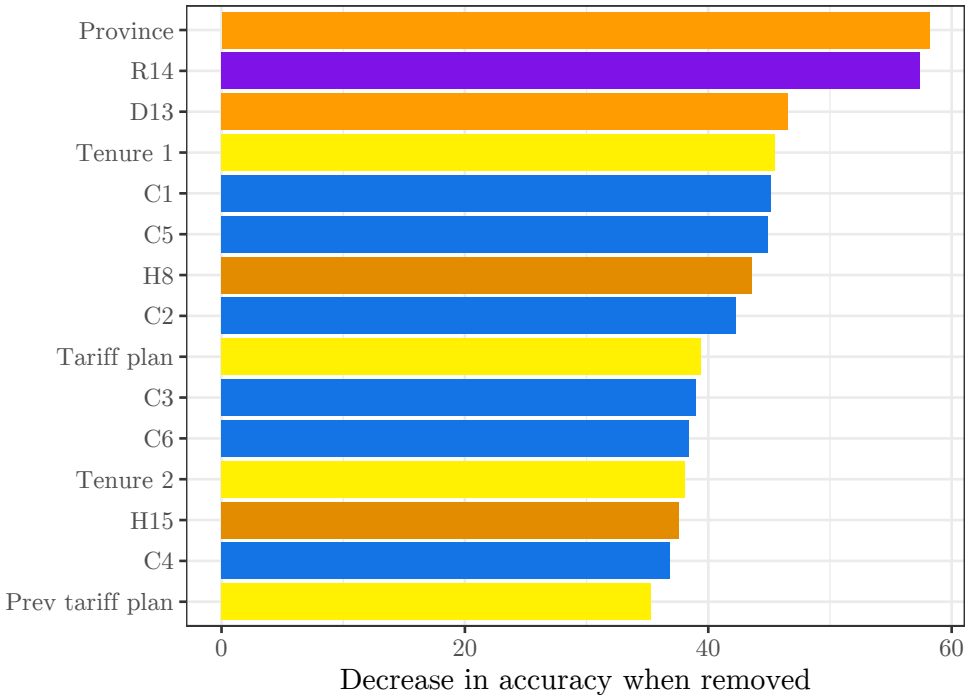


Figure 3.20 – Mean decrease in accuracy when a feature is removed for SIM only.

commitment of a customer to Orange, and the presence of this variable in both figures 3.23 and 3.24 expresses that it has also a monotonic relation to churn probability.

On the other hand, all others variables in this sensitivity analysis display a non-linear relationship to predicted churn probability. This is remarkably illustrated by the prominence of variables related to revenues on figure 3.23, colored in purple. These variables increase the predicted probability of churn when they are increased, as expected by the bill shock effect. However, they are absent from figure 3.24, indicating that a reduced bill is not associated to a reduced churn. It is worth mentioning that the age is associated in both graphs to an increased risk of churn. Consequently, any age far from the average age is associated to an increased risk of churn.

Bear in mind that this analysis is solely indicating statistical association between the values of different variables and the predicted probability of churn. It is by no mean an indication of causality. For example, the number of contracts is inversely associated to the risk of churn. It is tempting to conclude that selling new contracts to customers will therefore reduce their risk of stopping their subscription. Nevertheless, the analysis does not confirm this hypothesis: it might be the case that a satisfied customer is typically not likely not churn and is also more prone to buying new services. In this case, churn and number of contracts have a common cause (the customer satisfaction), and manipulating the number of contract will not modify the risk of churn. Also, the magnitudes of the difference in churn rates should not be considered as realistic, meaningful values. We added a standard deviation to each variable, regardless of whether a standard deviation make sense for each variable. For example, the number of contracts is typically not distributed according to a Gaussian distribution, since it takes as value only small positive integers. We can expect most of the decision trees

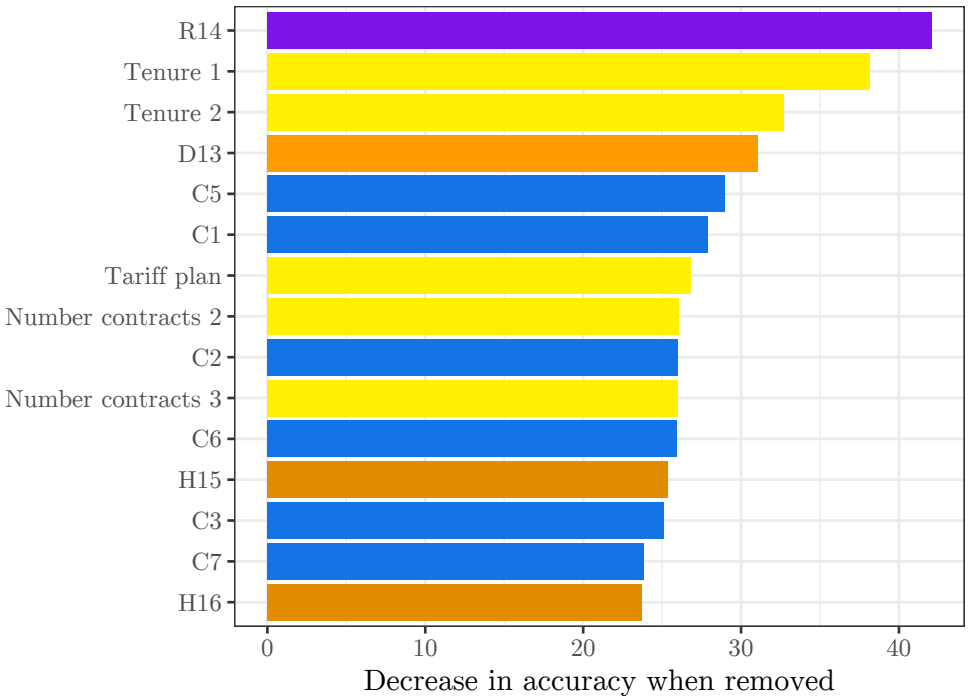


Figure 3.21 – Mean decrease in accuracy when a feature is removed for SIM only with difference and ratio variables.

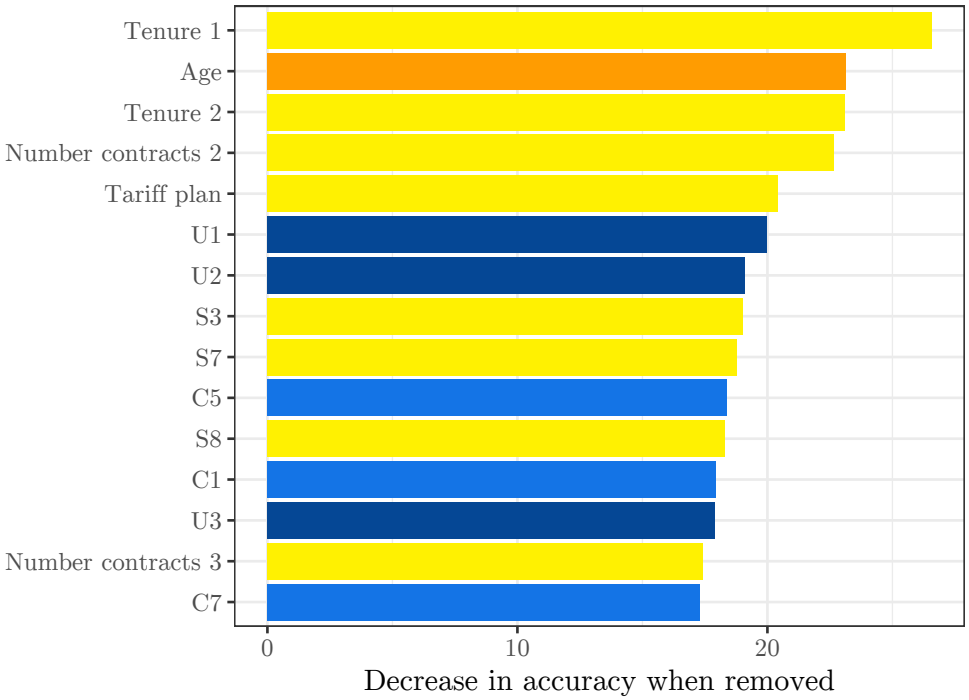


Figure 3.22 – Mean decrease in accuracy when a feature is removed for loyalty.

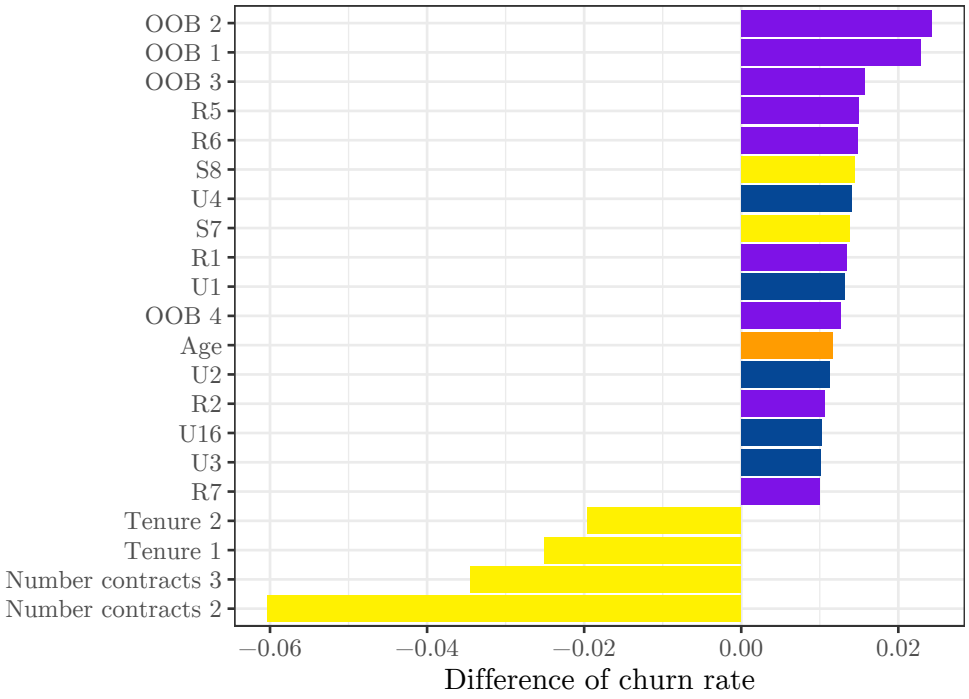


Figure 3.23 – Difference in predicted probability of churn when a standard deviation is added separately to each variable. Run on the SIM only dataset. Only variables inducing a difference having an absolute value greater than 0.01 are shown.

composing our model to choose a split point close to 1 for this variable, in order to categorize clients as either having 1 contract or more. Adding a standard deviation to the variable would change the path of each sample in these trees. This explains the disproportionate 15% of difference in churn rate in figure 3.24.

3.5 Comparison to state of the art

In this section we compare our results to other studies in churn prediction. Of the 20 articles in our bibliography related to churn prediction, 15 are empirical studies either suggesting a new method or comparing existing methods. 7 of these 15 articles use precision, recall, accuracy and F-measure as evaluation measures. These evaluation measures are applicable when the output of the prediction model is a hard label, such as for a support vector machine or a decision tree. However, our experiment uses an ensemble model composed of random forests, and the predictions take the form of a score between 0 and 1. This implies that a decision threshold has to be chosen when classifying customers as churners or non-churners. The precision, recall, F-measure and accuracy thus are functions of this threshold, and this does not allow a direct comparison with these 7 empirical studies. We are left with 8 other studies which use either the lift a different threshold (most often 10%, also named top decile lift), the expected maximum profit and the area under the ROC curve (AUROC). The results of these studies are compiled in table 3.3, along with our results in the last row.

We consider our results on the SIM only test set, taking the results of one of best

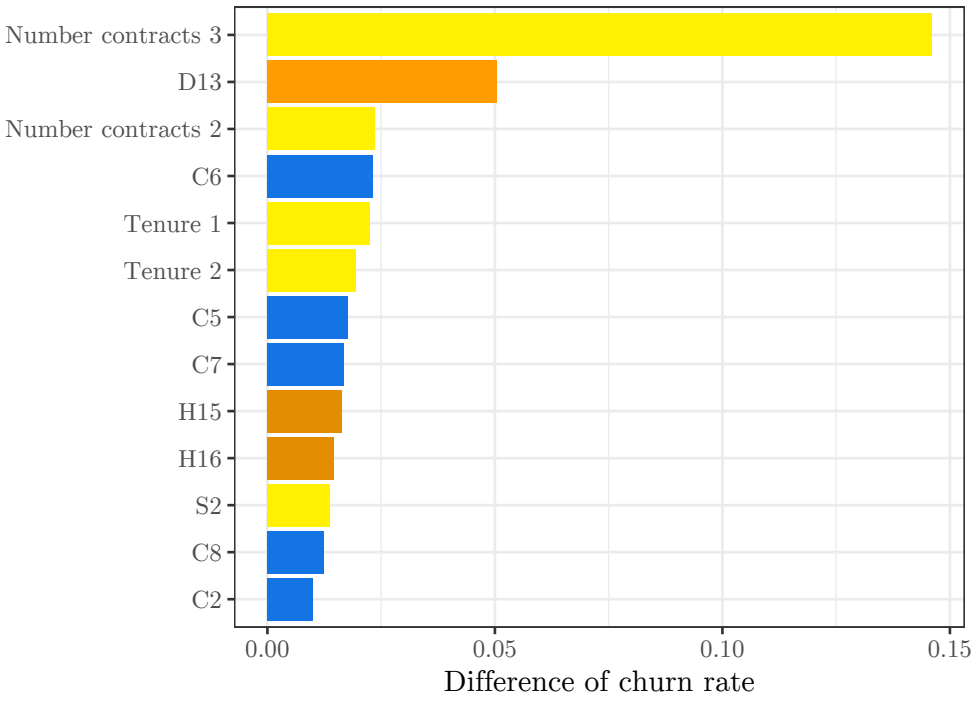


Figure 3.24 – Difference in predicted probability of churn when a standard deviation is subtracted separately from each variable. Run on the SIM only dataset. Only variables inducing a difference having an absolute value greater than 0.005 are shown.

Paper	Best method	AUROC	Lift 10%	Lift 5%
Coussement et al., 2017	Logistic regression	0.63	2.19	—
De Caigny et al., 2018	Logit Leaf model	0.87	5.34	—
Óskarsdóttir et al., 2018	Similarity forests	0.87	6.05	—
Zhu et al., 2017	C4.5 with UnderBagging	0.80	4.54	—
Mitrović et al., 2018	Random forest	0.74	2.35	—
Idris et al., 2014	Ensemble with mRMR	0.75	—	—
Óskarsdóttir et al., 2017	Logistic Regression	0.89	—	6.16
Verbeke et al., 2014	Relational classifier	—	3.11	3.92
Our results	Easy Ensemble	0.73	3.41	4.68

Table 3.3 – Comparison of our results to other studies in churn prediction.

performing configurations (no variable selection, no difference or ratio variables). It is important to notice that the studies mentioned in table 3.3 obviously do not provide a unique numerical result. In each of these studies, we considered, whenever possible, the results pertaining to a dataset similar to ours in terms of churn rate and type of contracts. Also, when multiple methods are compared in a study, we retained the evaluation measure of the method performing best. The name of the method retained in each study is given in the second column of the table.

In terms of area under the ROC curve, we achieve results similar to 2 studies, both using random forests (the ensemble proposed by Idris and Khan (2014) contains a random forest, a KNN, and a rotation forest). Also, we perform better than the logistic regression proposed by Coussement et al. (2017), but the 4 remaining papers outperforms our model by a clear margin. In terms of lift, we outperform the logistic regression in the first row, the random forest used in Mitrović et al. (2018) and the combined relational classifier proposed by Verbeke et al. (2014). All other studies yield a superior lift, both at 5% and 10% threshold.

3.6 Conclusion

We summarize here the main findings of our experiments on churn prediction.

- Feature selection does not reduce performance if at least 30 of the most important variables are selected.
- Adding difference and ratio variables reduces the performance if no feature selection is conducted beforehand.
- Due to a lucky domain shift, the trained models actually perform better on the test set than on the validation set.
- Churn is slightly easier to predict in the loyalty dataset, due to the importance of time-related variables.
- Important variables include, non-exhaustively: the tenure, the province, the tariff plan, the number of calls, and the data usage.
- The tenure and the number of contract are associated monotonically to the churn probability
- Variables related to the amount paid by the customer are associated to more churn when they are increased, but the opposite is not true.

Chapter 4

Causal analysis

4.1 Introduction

In this chapter, the application of causal inference to customer data is explored, in the hope of shedding light on the reasons for customer churn. A predictive experiment as conducted in the previous chapter indicates which variables are indicative of a client about to churn, but there is no guarantee that an intervention on any of these variables will have a positive effect. For example, the number of phone calls has a strong predictive power as shown in figure 3.20. However, a hypothetical churn retention action that would make customers call less often will maybe fail, if clients about to churn tend to call less because they are unsatisfied of the service. In this case, the predictive variable is an effect of churn and modifying it has no causal impact. Different tools are needed to discover true causal relationships between variables. In this chapter, we focus on three types of models: causal Bayesian networks, information-theoretic filters, and supervised causal inference. An overview of state-of-the-art methods for each of these models is given in section 2.2, but we describe them here in more details.

We begin by introducing *Bayesian networks*, which are graphical models used to represent probabilistic dependencies between random variables. They are represented by a *directed acyclic graph* (DAG) where the nodes are random variables, and a joint probability density is assigned to these variables. We will use the terms *nodes* and *variable* interchangeably. In a directed acyclic graph, a node A is a parent of B if there is a direct edge from A to B, A is an ancestor of B if there is a direct path from A to B. We can define in a similar way the notions of child, descendant, non-descendant and spouse in a directed acyclic graph. Bayesian networks come in a causal variant, which is defined here (Guyon, Aliferis, et al., 2007).

Definition 1 (Causal Bayesian network)**.** Let $\boldsymbol{X}$ be a set of random variables and P a joint probability density over $\boldsymbol{X}$. Let Γ be a DAG in which the vertices are $\boldsymbol{X}$. It is required that

(i) for every edge from a node $X \in \boldsymbol{X}$ to a node $Y \in \boldsymbol{X}$, X is a direct cause of Y, and

(ii) for every node $X \in \boldsymbol{X}$, X is probabilistically independent of all its non-descendants, given its parents.

The first condition is required for a Bayesian network to be causal, and the second condition is called the Markov condition. The tuple $(\boldsymbol{X}, P, \Gamma)$ is a causal Bayesian network iff both conditions are satisfied.

We denote independence between two variables A and B according to the probability density P as $A \perp_P B$. Similarly, the conditional independence between two variables A and B given a set of variables $\boldsymbol{C}$ is written $A \perp_P B|\boldsymbol{C}$. Using the notion of *d-separation* as introduced by Pearl (*e.g.* in Pearl, 2002), we define the conditional independence between two variables A and B entailed by the Markov condition on a DAG Γ as $A \perp_\Gamma B|\boldsymbol{C}$.

The condition of *faithfulness* is often set onto a causal Bayesian network:

Definition 2 (Faithfulness)**.** A DAG Γ is *faithful* to a joint probability density P iff every dependency entailed by the Markov condition on Γ is also entailed by P. That is,

$$\forall X, Y \in \boldsymbol{X}, \forall \boldsymbol{Z} \subset \boldsymbol{X}, \quad X \not\perp_\Gamma Y|\boldsymbol{Z} \Rightarrow X \not\perp_P Y|\boldsymbol{Z}. \tag{4.1}$$

Both the Markov conditions and the faithfulness conditions ensure that a given graph and a given probability density represent accurately the same set of dependencies and independencies. When both conditions are met, we write (in)dependence relations without specifying whether it is entailed by P or Γ. The faithfulness condition, in particular, ensures that the influence of a cause onto an effect by multiple causal routes does not cancel itself out. To demonstrate a violation of this assumption, assumes that a gene codes both for the production of a particular protein and for the suppression of another gene that codes this protein as well. In this case, when the first gene is removed, the protein is still produced by the other gene, and therefore the presence of the first gene seems to be independent of the production of the protein. The probability density P postulates the independence between the two variables, whereas the causal structure of the problem indicates a causal link(Hitchcock, 1997).

Another important concept in Bayesian networks is the notion of *Markov blanket* of a variable. Informally, this is a set of variables that shields a given target variable from the influence of the rest of the network. This notion is useful, for example, in feature selection for machine learning. The Markov blanket of a target variable is a subset that brings the maximum information about the target, among all possible sets of variables. Adding any other variable will bring some information that is already contained in other members, or in the interaction thereof, of the Markov blanket.

Definition 3 (Markov blanket)**.** For a set of random variables $\boldsymbol{X}$, a subset $\boldsymbol{M} \subset \boldsymbol{X}$ and a variable $Y \in \boldsymbol{X}$, $\boldsymbol{M}$ is the Markov blanket of Y iff for any subset $\boldsymbol{V} \subset \boldsymbol{X}$, Y is conditionnally independent of $\boldsymbol{V} \setminus \boldsymbol{M}$ given $\boldsymbol{M}$.

In our case, where causal Bayesian networks are represented by a directed acyclic graph, and where both the Markov and the faithfulness conditions are met, the Markov blanket of a variable is unique and consists of its direct causes, its direct effects, and the direct causes of its direct1235711 effects. An example is given on figure 4.3.

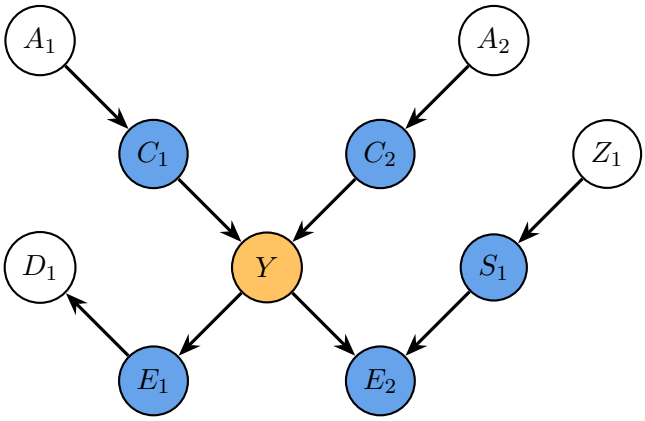


Figure 4.1 – A causal Bayesian network, with the Markov blanket of Y highlighted in blue.

These definitions lay the theoretical background for graphical causal models, but methods for inference from observational data still need to be derived. The methods used in this experiment are described in section 4.4. A common requirement of some of these methods is an accurate test of statistical independence between two variables in $\boldsymbol{X}$. Many statistical independence tests exist, but in our case, we need an estimator that is able to handle a mix of categorical and continuous variables. We use the *mutual information*, which is an information-theoretic measure of statistical dependency. It is more general than the Pearson or the Spearman correlation coefficient, as it encompasses any type of dependency, and not only linear or monotonic relationships. Also, it is defined for any two random variables, regardless of their type or domain. In the case of two discrete variables X and Y taking values respectively in $\mathcal{X}$ and $\mathcal{Y}$, with a joint probability distribution $P(X,Y)$ and marginal probability distributions $P(X)$ and $P(Y)$, the mutual information between X and Y is defined as (Cover & Thomas, 2012)

$$I(X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right) \tag{4.2}$$

$$= H(X) - H(X|Y) \tag{4.3}$$

$$= H(Y) - H(Y|X) \tag{4.4}$$

$$= H(X) + H(Y) - H(X,Y) \tag{4.5}$$

where $H(X)$ is the entropy of X and $H(X|Y)$ is the conditional entropy of X given Y(Shannon, 1948). The two last equalities indicate that the mutual information is a symmetric, positive quantity, and that it can be viewed as the reduction in uncertainty that a variable brings about the other. A schematic view of these formulae is given in figure 4.2. In the case of two continuous variables, an analogous definition exist (Kolmogorov, 1956)

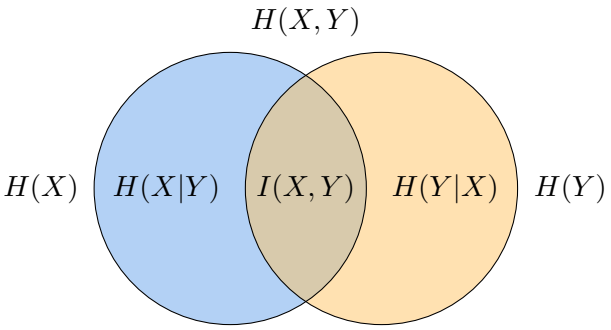


Figure 4.2 – Schematic representation of the relationship between entropy, conditional entropy, joint entropy, and mutual information of two discrete random variables.

$$I(X,Y) = \int_{\mathcal{X}}\int_{\mathcal{Y}} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)dx\,dy \tag{4.6}$$
$$= h(X) - h(X|Y) \tag{4.7}$$
$$= h(Y) - h(Y|X) \tag{4.8}$$
$$= h(X) + h(Y) - h(X,Y) \tag{4.9}$$

where $h(X)$ is the differential entropy of X and $h(X|Y)$ is the conditional differential entropy of X given Y. In the case of two normally distributed variables, we have

$$I(X,Y) = -\frac{1}{2}\log(1-\rho^2) \tag{4.10}$$

where ρ is the Pearson correlation coefficient between X and Y. Even though the assumption of normal distribution does not hold in general, empirical results show that it is still a decent estimator for non-linear dependencies (Olsen, Meyer, & Bontempi, 2008). Finally, the mutual information between a continuous variable X and a discrete variable Y taking values in a finite set $\mathcal{Y}$ can be computed as

$$I(X,Y) = h(X) - h(X|Y) = h(X) - \sum_{y\in\mathcal{Y}} h(X|Y=y)P(Y=y) \tag{4.11}$$

which means that the mutual information in the mixed case can be computed with an estimator of the differential entropy of a continuous variable. Following the discretization method described in (Olsen et al., 2008), let N be the number of values sampled iid from X. We divide the domain $\mathcal{X}$ of X into k bins of equal size Δ, and we write $\mathrm{nb}(i)$ the number of samples present in the ith bin, $\forall i \in \{1,\dots,k\}$. The differential entropy of X is estimated with the *Miller-Madow estimator*

$$\hat{h}(X) = \sum_{i=1}^{k}\frac{\mathrm{nb}(i)}{N}\log\left(\frac{\mathrm{nb}(i)}{N}\right) + \log\Delta + \frac{k-1}{2N} \tag{4.12}$$

It follows that we can compute the mutual information in all possible configurations of variable types:

(i) Between two discrete variables, using equation 4.2

(ii) Between a continuous variable X and a discrete variable Y with equation 4.11 and the differential entropy estimator of equation 4.12

(iii) Between two continuous variables assuming normal distributions with equation 4.10

The mutual information can be generalized for any n variables, and is then named n-way *interaction* or *co-information* (A. J. Bell, 2003). A general formula exists, but we are only interested in the case $n = 3$ since it is connected with causal configurations composed of three variables. In this case, the 3-way interaction (that we simply name interaction) between three random variables X_1, X_2, and Y is defined in (McGill, 1954) as

$$I(X_1, X_2, Y) = I(X_1, X_2) - I(X_1, X_2|Y) \tag{4.13}$$

where, in the case of a discrete Y taking values in a finite set $\mathcal{Y}$, $I(X_1, X_2|Y)$ can be computed as

$$I(X_1, X_2|Y) = \sum_{y \in \mathcal{Y}} P(Y = y) I(X_1, X_2|Y = y) \tag{4.14}$$

A more general definition of conditional mutual information is given in (Cover & Thomas, 2012). But since we consider a classification problem with $\mathcal{Y} = \{0, 1\}$, this restricted definition is sufficient for our purposes.

4.2 Scope

In this section, the same dataset as in the predictive modeling part is used. We restrict ourselves to SIM only contracts since it is supposed that the causes of churn are at least partially different between loyalty and SIM only contracts. All 5 months of data are used. In order to decrease computation time, only the first 30 variables in the ranking of the random forest trained in chapter 3 are used. Depending on the algorithm being used, a random subsampling has been applied, also to achieve decent computation times. In all cases, the positive class (churners) is untouched, and a random subset of the negative class is sampled so that the class ratio is even.

4.3 Prior knowledge on causes of churn

4.4 Experiments

The overall scheme of this experiment consists of running several causal inference techniques, which give different types of results in various forms, and extract a general consensus, if any, in the light of the different assumptions each model put on the data. Indeed, all causal inference methods are based on different assumptions, and the ability of a given method to infer causal patterns from observational data lies upon these assumptions.

More specifically, we use 5 different causal inference algorithms:

- PC
- Grow-shrink (GS)
- Incremental Association Markov Blanket (IAMB)
- Minimum interaction maximum relevance (mIMR)
- D2C

For the first three methods, we use the R package *bnlearn* (Scutari, 2009) for independence tests using mutual information and asymptotic χ^2 test (Good, 2013). For mIMR and D2C, we use the R package *D2C* (Bontempi & Flauder, 2015), along with another implementation of mIMR using the mutual information estimator given in section 4.1. The sample size used in the experiment is given after the description of each algorithm. In all cases, a false positive rate of 0.05 is chosen for statistical tests of independence.

PC

Description The PC algorithm (Spirtes & Glymour, 1991) returns the set of directed acyclic graphs that are faithful to a given probability distribution. It is based on independence tests between two variables, conditioned on a set of other variables. It uses the notion of d-separation to eliminate or find the direction of putative causal links. The PC algorithm is given in algorithm 1, where **Adjacencies**(X) is the current set of nodes that are adjacent to X in Γ. Therefore, it evolves as Γ is modified in the algorithm. The idea underlying PC is to A) start with a full graph, B) remove edges using independence tests with conditioning sets of increasing size, C) orient colliders using the d-separation property, and D) find remaining orientations using two more rules. The assumptions underlying this algorithm are

(i) There is no unmeasured confounder

(ii) The statistical tests are correct

(iii) The causal relationships between variables are the same for all samples

(iv) There exists a DAG Γ representing the causal structure that is faithful to the underlying joint probability density P (definition 2)

If these assumptions hold, then the result of the algorithm is a set of DAGs that are all faithful to the density probability P. The assumption (iii) is reasonable in our case, but the three others less so.

Experimental setting The PC algorithm is slow when the number of samples is large since the whole Bayesian network is inferred. Therefore, we restrict the dataset to 10,000 samples. The implementation given in the package *bnlearn* is used. The results are given under the form of a directed acyclic graph.

Algorithm 1 The PC algorithm

A) Let Γ be a complete undirected graph on the set of vertices $\boldsymbol{X}$.
B) $n \leftarrow 0$
repeat
  repeat
    Select a pair of vertices X and Y such that
      • X and Y are adjacent
      • there exists a set $\boldsymbol{Z} \subseteq$ **Adjacencies**$(X) \setminus \{Y\}$ that has n elements
      • $X \perp Y | \boldsymbol{Z}$.
    Remove the edge X—Y from Γ
    Add $\boldsymbol{Z}$ to **Sepset**(X, Y) and to **Sepset**(Y, X)
  until no X, Y pairs satisfying above conditions can be found
  $n \leftarrow n + 1$
until for each pair X and Y, $|$**Adjacencies**$(X) \setminus \{Y\}| < n$
C) For all triplets X—Y—Z where there is no edge between X and Z, orient it as $X \rightarrow Y \leftarrow Z$ if Y is not in **Sepset**(X, Z)
D)
repeat
  Orient all $X \rightarrow Y$—Z as $X \rightarrow Y \leftarrow Z$
  Orient all X—Y as $X \rightarrow Y$ if there is a directed path from X to Y
until no more edges can be oriented

Grow-Shrink

Description The GS algorithm (Margaritis & Thrun, 2000) is a Markov Blanket discovery algorithm efficient even for a large number of variables. It is based on an estimator that returns a numerical value for the statistical dependency between two variables, potentially conditioned on a set of other variables. In our case, we use the mutual information. Consider a target variable Y and a set of predictor variables $\boldsymbol{X}$. The GS algorithm constructs the Markov blanket of Y, denoted $\mathbf{MB}(Y)$, in three phases performed in sequence:

A) All variables $X \in \boldsymbol{X}$ are ordered in decreasing order according to $I(X, Y)$

B) Each variable $X \in \boldsymbol{X}$ is added to $\mathbf{MB}(Y)$ iff it is conditionally dependent on Y, given $\mathbf{MB}(Y)$. That is, iff $I(X, Y|\mathbf{MB}(Y)) > 0$.

C) Each variable $X \in \mathbf{MB}(Y)$ is removed from $\mathbf{MB}(Y)$ iff it is conditionally independent of the rest of $\mathbf{MB}(Y)$. That is, iff $I(X, Y|\mathbf{MB}(Y) \setminus \{X\}) = 0$.

The phase A) is a heuristic to speed up the search, however Tsamardinos, Aliferis, Statnikov, and Statnikov (2003) pointed out that this delays the inclusion of spouses, since those have small unconditional relevance to Y. Therefore, more false positives are included before spouses get into $\mathbf{MB}(Y)$.

Experimental setting The entire set of positive samples is used, along with a subset of the same size of negative samples. This amounts to a total of 240,168 samples. The GS algorithm has been implemented using the independence test in the package *bnlearn*. The results are given as a list of members of the Markov blanket.

Incremental Association Markov Blanket

Description The IAMB algorithm (Tsamardinos, Aliferis, et al., 2003) is essentially similar to the GS algorithm, but does not include the sorting heuristic as a first phase:

A) The variable $X \in \boldsymbol{X}$ maximizing $I(X, Y|\mathbf{MB}(Y))$ is added to $\mathbf{MB}(Y)$, repeatedly until all remaining variables are independent of Y given $\mathbf{MB}(Y)$.

B) Each variable $X \in \mathbf{MB}(Y)$ is removed from $\mathbf{MB}(Y)$ iff it is conditionally independent of the rest of $\mathbf{MB}(Y)$. That is, iff $I(X, Y|\mathbf{MB}(Y) \setminus \{X\}) = 0$.

Experimental setting The experimental setting is the same as for the GS algorithm.

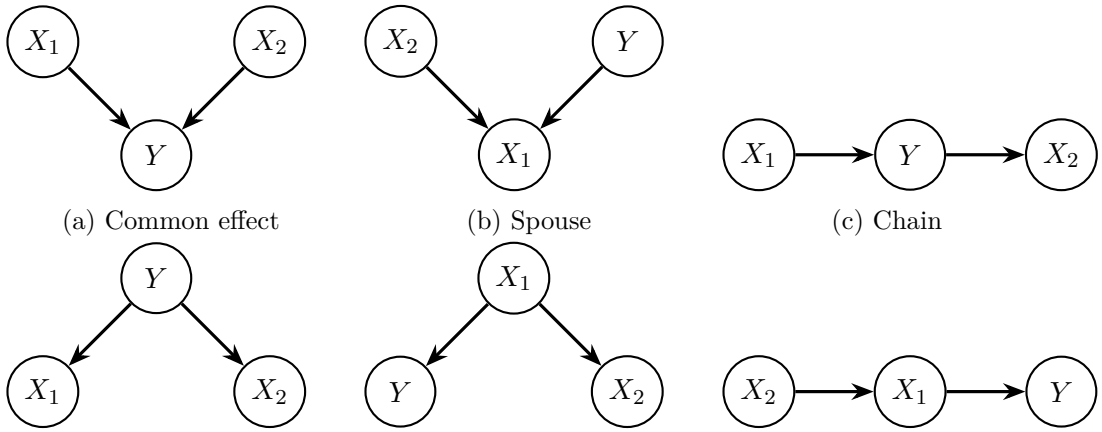
Minimum Interaction Maximum Relevance

Description The mIMR filter (Bontempi & Meyer, 2010) is a feature selection algorithm that has similarities with the mRMR algorithm (Peng et al., 2005). By using the interaction instead of the redundancy between the candidate variable and the set of selected variables, causes and spouses of the target are favored. In order to select only causes, spouses are eliminated beforehand on the basis of their null unconditional mutual information with the target. More formally, the main objective of feature selection is to find a subset $\boldsymbol{X}^*$ of a set of variables $\boldsymbol{X}$ that maximizes the mutual information with the target Y:

$$\boldsymbol{X}^* = \underset{\boldsymbol{X}_S \subseteq \boldsymbol{X}}{\arg\max}\, I(\boldsymbol{X}_S, Y) \tag{4.15}$$

Evaluating all possible subsets $\boldsymbol{X}_S$ is computationally infeasible, the forward selection scheme is therefore adopted. It consists in repetitively selecting the variable X^*_{d+1} that bring the most improvement given the set $\boldsymbol{X}_S$ containing the d variables already selected, until a fixed number v of variables is attained. The improvement is determined in a way that favors direct causes of Y. Consider the interaction equation 4.13. The interaction $I(X_1, X_2, Y)$ can be viewed as the reduction in statistical dependency between X_1 and X_2 brought by the knowledge of Y. On the one hand, a positive interaction occurs in 4 types of causal patterns, shown in figures 4.3c, d, e and f. On the other hand, a negative interaction occurs only in the common cause configuration (figure 4.3d) and the spouse configuration (figure 4.3b). Furthermore, one can differentiate between common causes and spouses by noticing that a spouse of the target has null unconditional relevance with the target (a spouse is relevant only when the common effect is known). In practice, a statistical test is used to determine the set of variables having non-null mutual information with the target, written $\boldsymbol{X}_+$. This leads to the update criterion of mIMR that is a linear combination of relevance and interaction:

$$X^*_{d+1} = \underset{X_k \in \boldsymbol{X}_+ \setminus \boldsymbol{X}_S}{\arg\max}\, \left[I(X_k, Y) - I(\boldsymbol{X}_S, X_k, Y) \right]. \tag{4.16}$$



(a) Common effect (b) Spouse (c) Chain

(d) Common cause (e) Brotherhood (f) Grandparent

Figure 4.3 – Possible causal patterns between two variables X_1 and X_2 and a target variable Y having non-null interaction.

Using the approximation

$$I(\boldsymbol{X}_S, X_k, Y) \approx \frac{1}{d} \sum_{X_i \in \boldsymbol{X}_S} I(X_i, X_k, Y), \tag{4.17}$$

the forward step can be written as

$$X^*_{d+1} = \underset{X_k \in \boldsymbol{X}_+ \setminus \boldsymbol{X}_S}{\arg\max} \left[I(X_k, Y) - \frac{1}{d} \sum_{X_i \in \boldsymbol{X}_S} I(X_i, X_k, Y) \right] \tag{4.18}$$

$$= \underset{X_k \in \boldsymbol{X}_+ \setminus \boldsymbol{X}_S}{\arg\max} \left[I(X_k, Y) - \frac{1}{d} \sum_{X_i \in \boldsymbol{X}_S} \left(I(X_i, X_k) - I(X_i, X_k, |Y) \right) \right] \tag{4.19}$$

$$= \underset{X_k \in \boldsymbol{X}_+ \setminus \boldsymbol{X}_S}{\arg\max} \left[I(X_k, Y) + \frac{1}{d} \sum_{X_i \in \boldsymbol{X}_S} \sum_{y \in \mathcal{Y}} P(Y = y) \left(I(X_i, X_k, |Y = y) - I(X_i, X_k) \right) \right] \tag{4.20}$$

where equation 4.19 is derived using equation 4.13 and similarly 4.20 is derived from 4.14. The first two variables are selected as

$$X^*_1, X^*_2 = \underset{X_i, X_k \in \boldsymbol{X}_+}{\arg\max}\ I([X_i, X_k], Y). \tag{4.21}$$

The mIMR filter is based on some underlying estimator of mutual information between two variables, and a statistical test of independence for selecting the set of unconditionally relevant variables.

Experimental setting In this experiment, two implementations are used:

- One using the mutual information estimator described in section 4.1 and the test of independence in the package *bnlearn*.

- One assuming normally-distributed continuous variables, allowing to compute the mutual information (with equation 4.10) and to test for independence using the Pearson correlation coefficient. Discrete variables are converted to numerical values using one-hot encoding.

The drawback of the first method is that the mutual information estimator is ad-hoc: it assumes a monotonic relationship between two continuous variables but uses a histogram-based entropy estimator in the mixed case. This may lead to inconsistencies in the measure of mutual information. On the other hand, the second method sets the linear assumption on all variables, even on one-hot encoded categorical variables. For the first implementation, the dataset is restricted to 10,000 samples, due to the computational cost of the entropy estimator. In the second implementation, 100,000 samples are used. The results are provided as a list of the first 15 selected variables, accompanied with the gain provided by each variable at each iteration of the algorithm.

D2C

Description The first three causal inference algorithms used in this section are solely based on statistical independence tests, and therefore are unable to differentiate between indistinguishable causal patterns, such as the two variables configuration or the fully-connected three variables configuration. Since the probability density P is reduced to a set of (in)dependence relations, any fully connected graph is faithful to P in these two cases. Asymmetrical patterns exist however in the joint probability density of a cause and its effect, as demonstrated by the results of the Kaggle competition Cause-effect pairs (https://www.kaggle.com/c/cause-effect-pairs). The D2C algorithm (Bontempi & Flauder, 2015) is based on the asymmetry of descriptor extracted from the Markov blanket of two causally linked variables. Consider two random variables X_1 and X_2 such that X_1 is a cause of X_2, and their respective Markov blanket $\mathbf{MB}(X_1)$ and $\mathbf{MB}(X_2)$. This setting is pictured in figure 4.4. We consider only one cause, effect and spouse per Markov blanket for the sake of the presentation, but the principles generalize obviously to any Markov blanket. Even though we cannot distinguish between causes, effects and spouses among $\mathbf{MB}(X_1)$ and $\mathbf{MB}(X_2)$, we can derive several inequalities using d-separation. Consider the variables M_1 and M_2, which are members of respectively $\mathbf{MB}(X_1)$ and $\mathbf{MB}(X_2)$, but whose relation to X_1 and X_2 is unknown (that is, M_1 is either C_1, E_1 or S_1). We have

$$\begin{cases} I(X_1, M_2|X_2) > I(X_2, M_1|X_1) & \text{if } M_2 = C_2 \\ I(X_1, M_2|X_2) = I(X_2, M_1|X_1) & \text{otherwise,} \end{cases} \tag{4.22}$$

since the only collider configuration between one of X_1 and X_2 and a member of the Markov blanket of the other variable is $X_1 \to X_2 \leftarrow C_1$. By computing a population of descriptors

$$D(1,2) = \{I(X_1, M_2|X_2)|\ \forall M_2 \in \mathbf{MB}(X_2)\} \tag{4.23}$$

$$D(2,1) = \{I(X_2, M_1|X_1)|\ \forall M_1 \in \mathbf{MB}(X_1),\}, \tag{4.24}$$

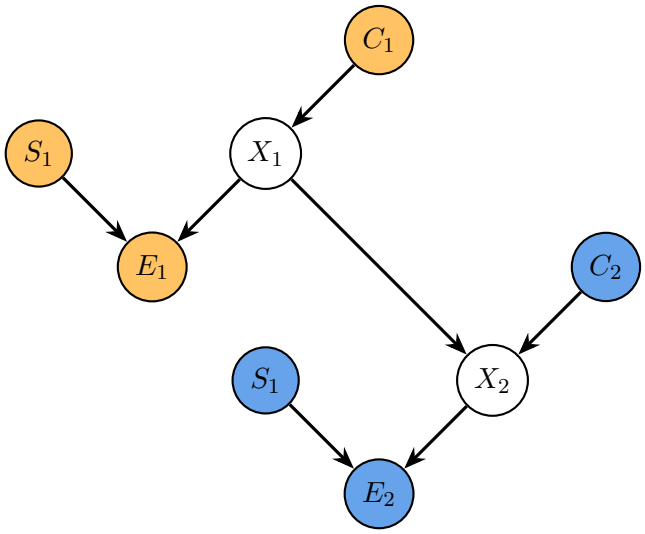


Figure 4.4 – Two causally linked variables and their Markov blanket.

equation 4.22 indicates that the distribution of $D(1,2)$ differs from $D(2,1)$. Other similar inequalities are used to compute 2 other population of descriptors. The rank of each M_1 in $\mathbf{MB}(X_2)$ is also computed, along with the rank of each M_2 in $\mathbf{MB}(X_1)$. The quartiles of the population of these various descriptors are computed, along with the mutual information between X_1 and X_2, and the mutual information conditioned on $\mathbf{MB}(X_1)$ or $\mathbf{MB}(X_2)$. All these quantities are then used as features on a machine learning algorithm, whose task is to predict the probability of a causal link between X_1 and X_2. The default implementation of D2C uses a random forest classifier.

Experimental setting The D2C model is trained using randomly generated DAGs, as described in (Bontempi & Flauder, 2015) and implemented in the R package *D2C*. We use 50 DAGs having each a number of nodes sampled uniformly between 10 and 20, and each DAG generated 50 to 200 data samples. The function underlying the edge between two nodes is randomly chosen to be either linear, quadratic or a sigmoid. A Gaussian additive noise of standard deviation chosen randomly from 0.2 to 1 is added to each directed edge. The feature extraction phase uses the lazy learning approach (Bontempi, Birattari, & Bersini, 1999) to estimate mutual information, thus avoiding the linear assumption. We assume a Markov blanket of 4 variables when constructing the asymmetrical features. Given the high computational cost of feature extraction, 2,000 samples are used from the customer dataset. The results are provided as the predicted probability for each variable of being a cause of churn.

4.5 Results

In this section, the colors of the bars in graphs correspond to the variables categories presented in section 3.1:

- Subscription
- Calls and messages

- Mobile data usage
- Revenue
- Customer hardware
- Socio-demographic

PC

The output of the PC algorithm is a dense graph linking most of the variables, but unfortunately, the churn variable is completely disconnected from the rest of the graph. Note that it also the case for the province, the device manufacturer, and the tariff plan. All these variables are strongly informative for predicting churn but have been ruled out of the Markov blanket of the churn variable in this algorithm by conditional independence tests.

Grow-Shrink

The Markov blanket output by the GS algorithm contains 21 variables:

- 4 variables related to voice calls (C5, C6, C7, and C8)
- 6 variables related to data usage (U1, U2, U3, U4, U10, and U19)
- 4 variables related to messages (C1, C2, C3, C4)
- Out of bundle amount
- Age
- Tenure
- Number of contracts
- A socio-demographic variable (D13)
- A hardware-related variable (H15)
- A subscription-related variable (S7)

As explained by Tsamardinos, Aliferis, et al. (2003), the heuristic used in GS that favors variables having high unconditional relevance to the target increases the probability of false positive, since spouses of the target are not included in the beginning. Therefore, indirect causes or effects are included instead. This is particularly obvious in our case since the average duration of voice calls over the course of the three last months can be seen as a direct effect of the duration of voice calls in these three months. The average variable should therefore not be present in the Markov blanket.

Incremental Association Markov Blanket

The Markov blanket output by the IAMB algorithm contains 2 variables:

- The current tariff plan
- The previous tariff plan

According to IAMB, the churn is therefore independent of all other variables when we know the current and the previous tariff plan of the customer. This result is surprising but is stable for different values of false positive rates: a p-value of up to 0.2 was considered significant for independence tests, always giving the same result. It is noticeable that IAMB returns a Markov blanket comprising only categorical variables, while the GS algorithm gives only numerical variables. While the conditional independence tests are identical in the two methods, the different order of tests may have favored one type of variable over the other.

Minimum Interaction Maximum Relevance

Prior to the results of the mIMR algorithm, we show in figures 4.5 and 4.6 the mutual information and the interaction matrices of the 30 most predictive variables. These values are computed using our estimator presented in section 4.1. Let X_i and X_j be two variables in our dataset, and let Y be the churn variable. Figure 4.5 shows the mutual information $I(X_i, X_j)$ in the cell of position (i, j). A line and column are also added for the churn variable. On the one hand, one can see that no single variable seems to have a high mutual information with the churn. This explains why the PC and IAMB algorithms fail to find a satisfactory Markov blanket for the churn variable. On the other hand, 3 clusters of high mutual information can clearly be noticed, corresponding to the summary of voice calls, data usage and messages over the last 3 months. Given that these variables do not vary randomly from month to month, it is expected that they are strongly informative about each other, and even more about the average of the 3 months.

The matrix in figure 4.6 shows at position (i, j) the interaction between two variables X_i and X_j and the churn variable Y, that is, $I(X_i, X_j, Y)$. The row and column corresponding to the churn are not relevant in this figure. Recall that the interaction between two variables and a target is the reduction in statistical dependence that the knowledge of the target brings:

$$I(X_1, X_2, Y) = I(X_1, X_2) - I(X_1, X_2|Y).$$

It is also the amount of mutual dependence to the target that cannot be explained by bivariate interactions:

$$I(X_1, X_2, Y) = I(X_1, Y) + I(X_2, Y) - I([X_1, X_2], Y)$$

We thus seek couples of variables having a negative mutual information with the churn, since that means that those variables are complementary. Complementary variables are more likely to either be in a common effect or spouse configuration (figures 4.3a and b) with the churn. One couple stands out clearly in figure 4.6, the tenure and

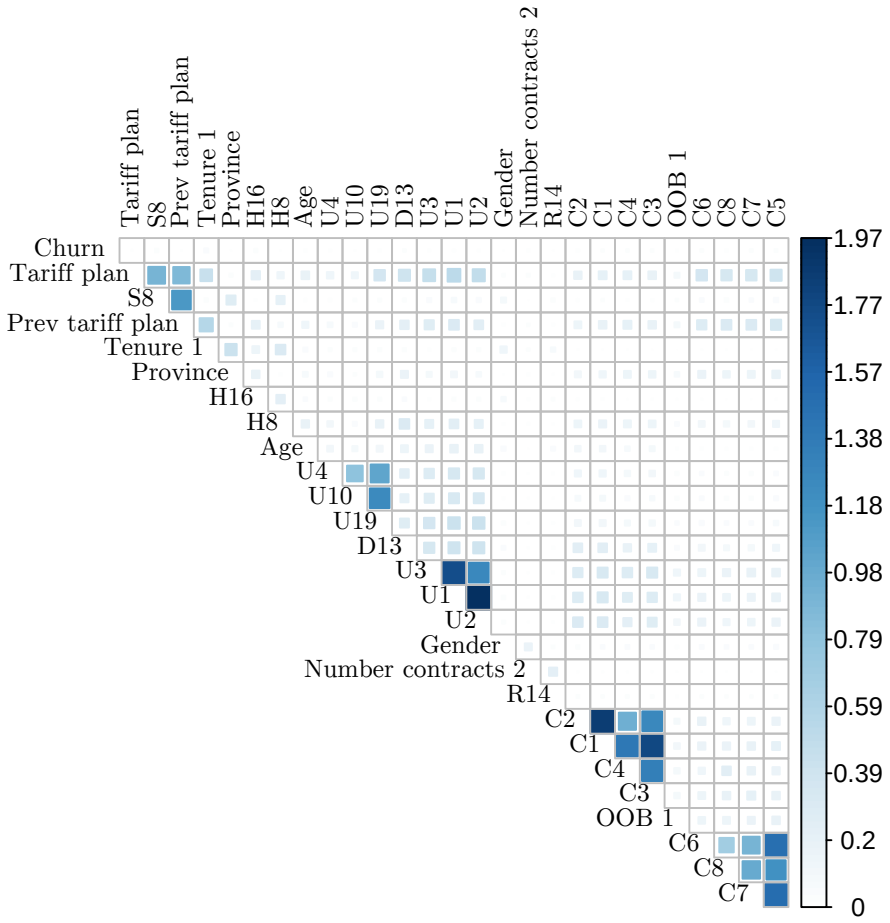


Figure 4.5 – Mutual information matrix

the province. The province negatively interacts with most other variables, meaning that it brings information about the churn only when considering it conjointly with other variables. The clusters of strongly correlated variables in figure 4.5 have a near-zero interaction since the knowledge of the churn does not change their distributions.

Figures 4.7 and 4.8 show the sequence of variables selected by the mIMR algorithm, for both our mutual information estimator (figure 4.7) and the estimator assuming Gaussian distributions (figure 4.8). Each row corresponds to one iteration of the algorithm. The width of the bar correspond to the value of the mIMR criterion at this step of the algorithm, that is, the approximated value of $I(X_k, Y) - I(\boldsymbol{X}_S, X_k, Y)$ where Y is the churn variable, X_k is the variable under consideration and $\boldsymbol{X}_S$ is the set of variables selected before X_k (i.e. above X_k on the plot). The two first variables have no gain, since they are directly selected as the pair of variables having the highest interaction with the target. Unsurprisingly, the first two variables in figure 4.7 are the tenure and the province (this couple of variable has the highest negative interaction in figure 4.6). Background knowledge, as well as figure 4.5, indicate that the following selected variables are not redundant with one another, up to the 9th and 10th rows. At these rows, the two variables are both related to the data usage. That probably indicates that the relevance term $I(X_k, Y)$ is prevailing over the interaction term $I(\boldsymbol{X}_S, X_k, Y)$.

The selected variables in figure 4.8 are mostly similar to those in 4.7, except that

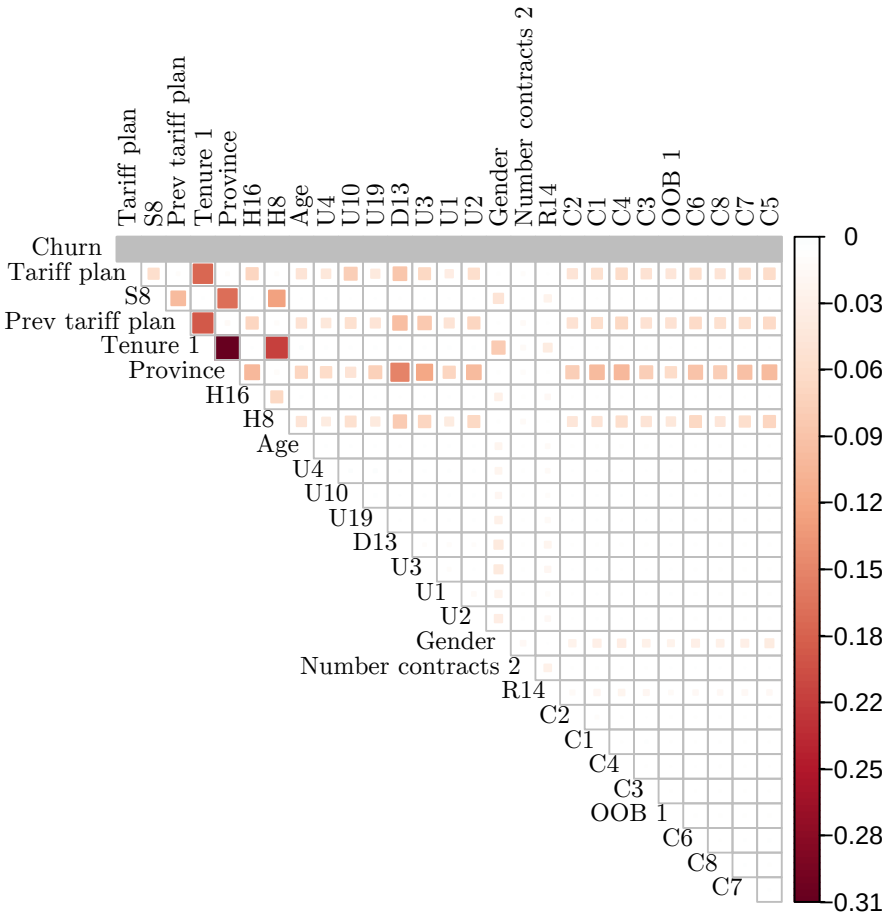


Figure 4.6 – Matrix of interaction between two variables and the churn.

all categorical variables are not in the first ranks. Since those are converted to as many numerical variables as there are categorical levels, the information is spread across multiple variables. Moreover, each of these new variables is considered to be Gaussian distributed, therefore not allowing to estimate optimally the mutual information. Another important difference between 4.7 and 4.8 is that the age and number of contracts are prevalent in the latter. The importance of age is most probably due to the Gaussian assumption, which is verified in this case, allowing efficient estimation of its mutual information with other variables.

D2C

The results from the D2C algorithm are shown in figure 4.9. To each variable correspond a probability of being a cause of churn predicted by the trained random forest. Since the implementation we use for D2C is designed for numerical variables, one-hot encoding is used. All the variables present in this plot are related to the tariff plan, previous tariff plan, province of residence and hardware-related variables. This is consistent with results of mIMR in figure 4.7, except for the second and last rows, related to the customer gender. Mobile data usage variables are also present and are the only variables related to service usage in this graph.

We do not believe that gender has a causal relationship with churn, and this rank

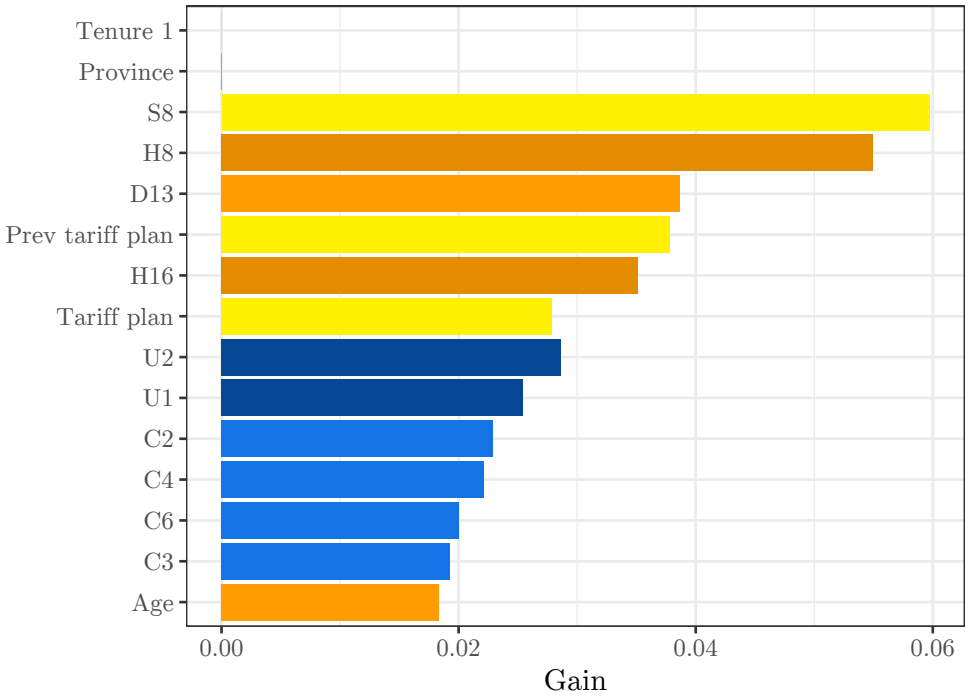


Figure 4.7 – Ranking of variables selected by mIMR with their respective gains, using the mutual information estimator of section 4.1. There is no gain for the first two variables.

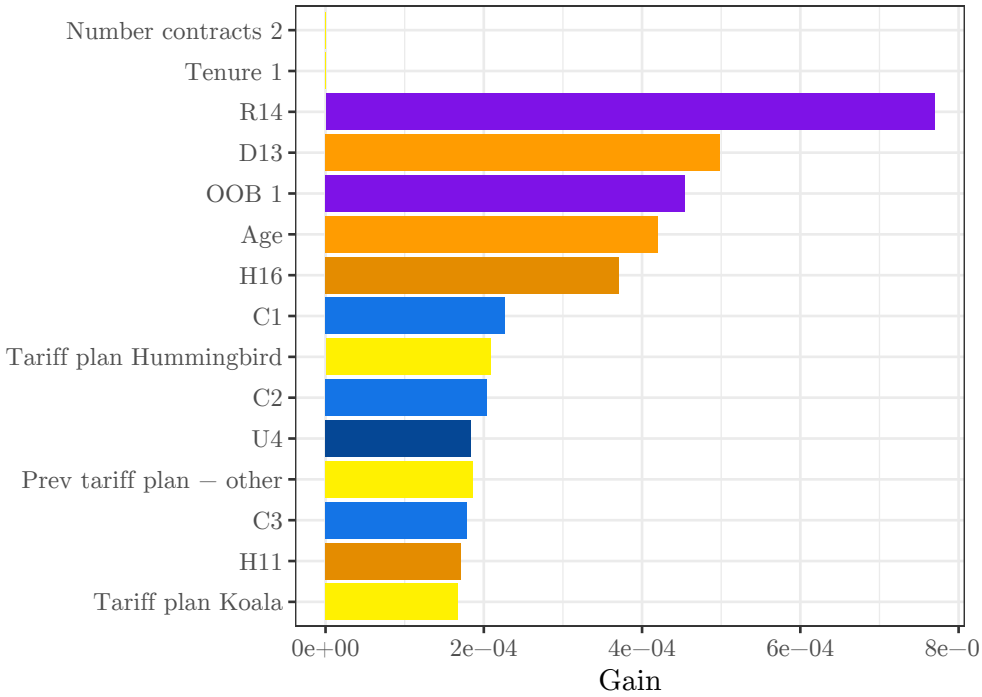


Figure 4.8 – Ranking of variables selected by mIMR with their respective gains, using one-hot encoding for categorical variables and assuming Gaussian distributions. There is no gain for the first two variables.

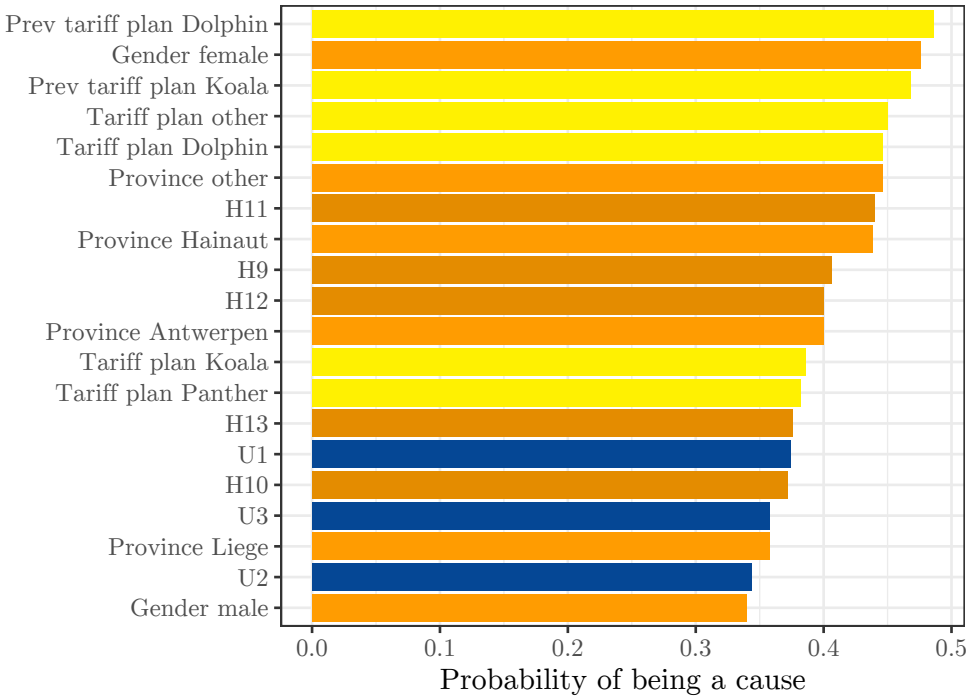


Figure 4.9 – Probability of causal link predicted by D2C

is most probably due to an artifact in the encoding of variables. Indeed, in the data preparation process, we noticed that data entries labeled as churners tend to have more often missing values in categorical variables. While this issue has been resolved, an unforeseen problem may have persisted in the gender variable.

4.6 Discussion

The output of the GS and IAMB algorithm correspond to the Markov blanket, indistinguishably causes, effects and spouses of churn. On the other hand, mIMR and D2C focus explicitly on direct causes, but a numerical score is provided for each variable. A choice of threshold has to be made on which variables we consider to be predicted as causes by these algorithms.

For the mIMR in figure 4.7, we include all variables up to the 9th and do not consider relevant the following ones. As discussed earlier, the 9th and 10th variables are mostly redundant with one another. Therefore, we can assume that at this threshold, the relevance term is becoming more important than the interaction term, and the causal property of the mIMR filter is not maintained anymore. In the case of mIMR with Gaussian assumption, the threshold is fixed at the 7th variable, since the following variables show a clear and distinct decrease in gain.

As for the output of the D2C algorithm, although there seems to be a large number of variables, most of them are different one-hot encodings of the same original variable. The most probable causes as predicted by D2C are solely the tariff plan, previous tariff plan, province of residence, device manufacturer and data usage. It seems reasonable to consider these 5 variables as predicted causes.

	PC	GS	IAMB	mIMR 1	mIMR 2	D2C
Tenure	✗	✔	✗	✔	✔	✗
Tariff plan	✗	✗	✔	✔	✗	✔
Prev tariff plan	✗	✗	✔	✔	✗	✔
Number contracts	✗	✔	✗	✗	✔	✗
Province	✗	✗	✗	✔	✗	✔
Age	✗	✗	✗	✗	✔	✗
Messages	✗	✔	✗	✗	✗	✗
Data usage	✗	✔	✗	✔	✗	✔
Voice calls	✗	✔	✗	✗	✗	✗
Out of bundle	✗	✔	✗	✗	✔	✗
H8	✗	✗	✗	✔	✗	✔
H14	✗	✔	✗	✔	✔	✗
S7	✗	✔	✗	✔	✗	✗
D13	✗	✔	✗	✔	✔	✗
R14	✗	✗	✗	✗	✔	✗

Table 4.1 – Summary of the results of causal analysis.

We summarize the results of the 5 algorithms (with the two implementations of mIMR) in table 4.1. For each variable, we indicate by which algorithm this variable was output, using the thresholds discussed above. There is no clear-cut consensus on which variables are causes. It is however reasonable to consider the number of messages and the duration of voice calls as *not* being inferred causes of churn, since only the GS algorithm outputs these variables. On the other hand, the tariff plan and previous tariff plan are given by all algorithms except for PC, GS, and mIMR with Gaussian assumption. We do not expect a categorical variable to be correctly predicted as a cause or not using the Gaussian assumption, the theory that the tariff plan and previous tariff plan are causes of churn are thus consistent with the observations.

In the absence of a formal procedure to assess these results, we could informally find arguments for or against considering each variable a cause of churn. This is due to the different assumptions laid by each method and the variety of their outputs. We will thus summarize the results of causal inference from observational data to the following statements:

- The messages, the voice calls, as well as all variables not represented in table 4.1, are considered as most probably *not causes of churn*.
- The tariff plan and previous tariff plan are considered as most probably *causes of churn*.
- The causal relationship of other variables in the table 4.1 is undecided.

Chapter 5

Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc ullamcorper sollicitudin dolor quisconvallis. Curabitur in dictum sapien. Suspendisse sed magna sodales, dictum lacus sed, porttitor neque. Duis semper vehicula elit, quis hendrerit ligula rutrum id. Praesent ac cursus leo, nec scelerisque eros. Aliquam imperdiet vestibulum turpis, ut hendrerit leo sollicitudin eu. Maecenas mattis elit quis lectus vestibulum rutrum. Nunc tempus metus sit amet sagittis porttitor. Integer varius metus aliquet purus pretium, a tincidunt justo faucibus. Mauris sollicitudin, tortor id tempus suscipit, odio diam dignissim dolor, quis rutrum libero nisl non nisl. Aliquam felis nisi, mollis nec finibus a, aliquam eu lectus.

Phasellus in ex hendrerit, malesuada tortor sit amet, sagittis enim. Phasellus at orci gravida ex malesuada congue ac ac felis. Cras maximus malesuada ligula id consequat. Praesent bibendum auctor massa eget dignissim. Ut sit amet rhoncus justo, ut pretium risus. Sed pellentesque nulla eget leo rhoncus malesuada. Vivamus at diam vitae nibh mattis malesuada non quis turpis. Etiam dignissim laoreet orci, non sollicitudin sem. Integer non sapien dui. Ut eu elit non turpis lobortis ullamcorper non a lacus. Integer dictum blandit fringilla. Quisque tempus metus in dolor rutrum, sed rutrum felis tristique. Nullam ut diam quis libero imperdiet facilisis. Aenean id nulla quis dolor maximus molestie. Donec et tortor vel ex hendrerit congue. Nullam pulvinar dui eu sollicitudin accumsan.

Aliquam nec lectus id leo sagittis feugiat consectetur id diam. Donec sed erat in turpis vehicula varius. Etiam pharetra dolor dolor. Praesent rhoncus dictum enim posuere rutrum. Cras et porttitor velit. Fusce condimentum turpis ut eros aliquet, et finibus dui facilisis. Nullam ullamcorper ex ac tortor dictum fermentum. Curabitur ultricies diam in consequat fringilla. Mauris a lectus dui. Mauris in est eget ligula fermentum vestibulum ac a odio. Aliquam erat volutpat. In ultrices purus id finibus volutpat. Phasellus elementum egestas pellentesque. Ut porttitor augue sed nunc pellentesque tempus. Vivamus tortor ante, posuere non magna at, gravida interdum arcu. Maecenas egestas blandit magna, sit amet tristique enim porta et.

Bibliography

Ahmed, A. A., & Maheswari, D. (2017). Churn prediction on huge telecom data using hybrid firefly based classification. *journal abbreviation*, *18*(3), 215–220. doi:10.1016/j.eij.2017.02.002

Aliferis, C. F. [Constantin F], Tsamardinos, I., & Statnikov, A. (2003). Hiton: A novel markov blanket algorithm for optimal variable selection. In *Amia annual symposium proceedings* (Vol. 2003, p. 21). American Medical Informatics Association.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20–29. doi:10.1145/1007730.1007735

Bell, A. J. (2003). The co-information lattice. In *Proceedings of the fifth international workshop on independent component analysis and blind signal separation: Ica* (Vol. 2003).

Bell, D. A., & Wang, H. (2000). A formalism for relevance and its application in feature subset selection. *Machine learning*, *41*(2), 175–195.

Bontempi, G., Birattari, M., & Bersini, H. (1999). Lazy learning for local modelling and control design. *International Journal of Control*, *72*(7-8), 643–658.

Bontempi, G., & Flauder, M. (2015). From dependency to causality: a machine learning approach. *The Journal of Machine Learning Research*, *16*(1), 2437–2457.

Bontempi, G., Haibe-Kains, B., Desmedt, C., Sotiriou, C., & Quackenbush, J. (2011). Multiple-input multiple-output causal strategies for gene selection. *BMC bioinformatics*, *12*(1), 458.

Bontempi, G., & Meyer, P. E. [Patrick E.]. (2010). Causal filter selection in microarray data. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 95–102).

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27–36. doi:10.1016/j.dss.2016.11.007

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Dahiya, K., & Bhatia, S. (2015). Customer churn analysis in telecom industry. In *2015 4th international conference on reliability, infocom technologies and optimization (icrito)(trends and future directions)* (pp. 1–6). IEEE.

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision

trees. *European Journal of Operational Research*, *269*(2), 760–772. doi:10.1016/j.ejor.2018.02.009

Fisher, R. A. (1937). *The design of experiments*. Oliver and Boyd; Edinburgh; London.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, *5*(Nov), 1531–1555.

Fonollosa, J. A. (2016). Conditional distribution variability measures for causality detection. *arXiv preprint arXiv:1601.06680*.

Good, P. (2013). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

Guyon, I., Aliferis, C. et al. (2007). Causal feature selection. In *Computational methods of feature selection* (pp. 75–97). doi:10.1201/9781584888796.ch4

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, *34*(10), 2902–2917.

Heckerman, D. (1998). A tutorial on learning with bayesian networks. In *Learning in graphical models* (pp. 301–354). Springer.

Hitchcock, C. (1997). Probabilistic causation.

Idris, A., & Khan, A. (2014). Ensemble based efficient churn prediction model for telecom. In *Frontiers of Information Technology (FIT), 2014 12th International Conference on* (pp. 238–244). doi:10.1109/fit.2014.52

Kayaalp, F. (2017). Review of customer churn analysis studies in telecommunications industry. *Karaelmas Fen ve Mühendislik Dergisi*, *7*(2), 696–705.

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, *24*, 994–1012. doi:10.1016/j.asoc.2014.08.041

Koller, D., & Sahami, M. (1996). Towards optimal feature selection (1996) proc. 13th int'l. conf. *Machine Learning*, 284–292.

Kolmogorov, A. (1956). On the shannon theory of information transmission in the case of continuous signals. *IRE Transactions on Information Theory*, *2*(4), 102–108.

Krzanowski, W. J., & Hand, D. J. (2009). *Roc curves for continuous data*. Chapman and Hall/CRC.

Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539–550. doi:10.1109/tsmcb.2008.2007853

Margaritis, D., & Thrun, S. (2000). Bayesian network induction via local neighborhoods. In *Advances in neural information processing systems* (pp. 505–511).

McGill, W. (1954). Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, *4*(4), 93–111.

Meyer, P. E. [Patrick Emmanuel], Schretter, C., & Bontempi, G. (2008). Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, *2*(3), 261–274.

Mitrović, S., Baesens, B., Lemahieu, W., & De Weerdt, J. (2018). On the operational efficiency of different feature types for telco Churn prediction. *European Journal of Operational Research*, *267*(3), 1141–1155. doi:10.1016/j.ejor.2017.12.015

Olsen, C., Meyer, P. E., & Bontempi, G. (2008). On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, *2009*(1), 308959.

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Systems with Applications*, *85*, 204–220. doi:10.1016/j.eswa.2017.05.028

Óskarsdóttir, M., Van Calster, T., Baesens, B., Lemahieu, W., & Vanthienen, J. (2018). Time series for early churn detection: Using similarity based classification for dynamic networks. *Expert Systems with Applications*, *106*, 55–65. doi:10.1016/j.eswa.2018.04.003

Pearl, J. (2002). Causality: Models, reasoning, and inference. *IIE Transactions*, *34*(6), 583–589.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (8), 1226–1238.

Sathe, S., & Aggarwal, C. C. (2017). Similarity Forests. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 395–403). KDD '17. doi:{10.1145/3097983.3098046}

Scutari, M. (2009). Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research*, *11*(May), 1643–1662.

Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, *9*(1), 62–72.

Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., & Snoeck, M. (2018). Profit maximizing logistic model for customer churn prediction using genetic algorithms. *Swarm and Evolutionary Computation*, *40*, 116–130. doi:10.1016/j.swevo.2017.10.010

Tsamardinos, I., Aliferis, C. F. [Constantin F.], & Statnikov, A. (2003). Time and sample efficient discovery of Markov blankets and direct causal relations. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 673–678). doi:10.1145/956804.956838

Tsamardinos, I., Aliferis, C. F. [Constantin F], Statnikov, A. R., & Statnikov, E. (2003). Algorithms for large scale markov blanket discovery. In *Flairs conference* (Vol. 2, pp. 376–380).

Umayaparvathi, V., & Iyakutti, K. (2016). Attribute selection and customer churn prediction in telecom industry. In *2016 international conference on data mining and advanced computing (sapience)* (pp. 84–90). IEEE.

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1–9. doi:10.1016/j.simpat.2015.03.003

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, *218*(1), 211–229.

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. *Applied Soft Computing*, *14*, 431–446. doi:10.1016/j.asoc.2013.09.017

Verbraken, T., Verbeke, W., & Baesens, B. (2013). A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE Transactions on Knowledge and Data Engineering*. doi:10.1109/tkde.2012.50

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research*, *5*(Oct), 1205–1224.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied artificial intelligence*, *17*(5-6), 375–381. doi:10.1080/713827180

Zhu, B., Baesens, B., & vanden Broucke, S. K. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information sciences*, *408*, 84–99. doi:10.1016/j.ins.2017.04.015