Understanding Telecom Customer Churn with Machine Learning: From Prediction to Causal Inference

Théo Verhelst[1](✉), Olivier Caelen[2], Jean-Christophe Dewitte[2], Bertrand Lebichot[1], and Gianluca Bontempi[1]

[1] Computer Science Department, Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium
{tverhels,blebicho,gbonte}@ulb.ac.be
[2] Data Science Team, Orange Belgium, Brussels, Belgium
{olivier.caelen,jean-christophe.dewitte}@orange.be

Abstract. Telecommunication companies are evolving in a highly competitive market where attracting new customers is much more expensive than retaining existing ones. Though retention campaigns may be used to prevent customer churn, their success depends on the availability of accurate prediction models. Churn prediction is notoriously a difficult problem because of the large amount of data, non-linearity, imbalance and low separability between the classes of churners and non-churners. In this paper, we discuss a real case of churn prediction based on Orange Belgium customer data. In the first part of the paper we focus on the design of an accurate prediction model. The large class imbalance between the two classes is handled with the EasyEnsemble algorithm using a random forest classifier. We assess also the impact of different data preprocessing techniques including feature selection and engineering. Results show that feature selection can be used to reduce computation time and memory requirements, though engineering variables does not necessarily improve performance. In the second part of the paper we explore the application of data-driven causal inference, which aims to infer causal relationships between variables from observational data. We conclude that the bill shock and the wrong tariff plan positioning are putative causes of churn. This is supported by the prior knowledge of experts at Orange Belgium. Finally, we present a novel method to evaluate, in terms of the direction and magnitude, the impact of causally relevant variables on churn, making the assumption of no confounding factors.

Keywords: Churn prediction · Machine learning · Big data · Causal inference

1 Introduction

In recent years, the number of mobile phone users had a massive increase, reaching more than 3 billion users worldwide. The number of mobile phone service subscriptions is greater than the number of residents in several countries, including

B. Bogaerts et al. (Eds.): BNAIC 2019/BENELEARN 2019, CCIS 1196, pp. 182–200, 2020.
https://doi.org/10.1007/978-3-030-65154-1_11

Belgium [13]. Telecommunication companies are evolving in a saturated market, where customers are exposed to competitive offers from many other companies. Hadden et al. [11] showed that attracting new customers can be up to six times more expensive than retaining existing ones. This led companies to switch from a sale-oriented to a customer-oriented marketing approach. By building customer relationships based on trustworthiness and commitment, a telecommunication company can reduce churn, therefore increasing benefits through the subsequent customer lifetime value. A typical marketing strategy to improve customer relationship is to conduct retention campaigns whose effectiveness depends on accurate profiles of customers (e.g. in terms of attrition risk).

Churn detection is nowadays performed by most major telecommunication companies using machine learning and data mining [5,12,18,28–31]. Churn prediction is a notoriously difficult learning task because of the large quantity of data, non-linearity, imbalance and low separability between the classes of churners and non-churners. The first part of the paper assesses several machine learning methods and strategies by using a large dataset measuring the churn behavior of Orange Belgium telecom clients. Estimating the probability of churn of a customer is however not sufficient if we wish to design an effective retention campaign (e.g. based on incentives). For this reason, the second part of the paper explores the adoption of causal techniques to infer from observational data the most probable causes of a churn behavior. Causal analysis is usually conducted through *controlled randomized experiments* [7], by evaluating the impact of a potentially causal variable on the target variable. In the context of customer relationship management, controlled experiments are possible through the retention campaigns, where the offers made to the customers act as variable manipulations. Though this reduces the risk of confounding factors, access to such data is typically difficult and expensive. For this reason, we have recourse to data-driven inference approaches, which aim to reconstruct causal dependencies based on the statistical distribution of the considered variables. Most existing approaches however make different assumptions about the data distributions which are difficult to assess in practice. For this reason, we adopt a "wisdom of the crowd" approach by running in parallel several state-of-the-art approaches and combining their results for final considerations. Also to assess the quality of the obtained putative causes we estimate from data the causal impact of every single cause on churn probability.

We may summarize the main contributions of the article as follows.

- Assessment of a state-of-the-art churn prediction pipeline and study of the impact of several model variants (e.g different feature sets and different subscription contracts) (Sect. 2).
- Application of causal strategies to infer putative causes of churn from observational data (Sect. 3).
- Assessment of the impact of putative causal variables on churn (Sect. 3).

The rest of this paper is structured as follows. In Sect. 2 we describe the dataset, the machine learning pipeline and the results of churn prediction. In Sect. 3 we provide a causal analysis of churn. Conclusion and future work perspectives are discussed in Sect. 4.

2 Churn Prediction

This section describes the Orange dataset and the machine learning pipeline designed to assess a number of strategies and models for predicting the probability of customer churn.

2.1 Data

The dataset is a monthly report of Orange Belgium customers' activity covering a 5 months time window in 2018. For confidentiality reasons, we will disclose here only some high-level details about the dataset. The dataset contains 73 features about customer activity including subscription type, used hardware, mobile data usage (in MB), number of calls/messages and some socio-demographic information. The dataset has 5.3 million entries (about 1 million entries per month). The target variable, denoting churn, is binary and takes the `true` value if the client is known to have churned in the two months following the input timestamp. The churn prediction problem is highly unbalanced, since there are far more non-churners than churners.

Two kinds of subscriptions are present in this dataset: SIM-only[1] and loyalty. The first refers to a subscription where the customer can quit at any time with no cost. In the loyalty contract the customer is rewarded (e.g. discount on the purchase of a mobile phone) in exchange of a fixed contract duration (e.g. 24 months). If the customer decides nonetheless to stop his subscription before the term of the contract, he has to pay back the reward. There are about 5 million entries in the SIM-only dataset, and about 250,000 entries in the loyalty dataset. In this paper, we will mainly focus on SIM-only contracts, given its broader impact on the Orange customer base and the larger statistical power due to the availability of more samples. Some experiments have been conducted anyway on both contract types, to understand the differences in terms of churn behavior.

In order to provide a visual description of the informative content of the dataset, let us consider in Fig. 1 two variables having a clear relation with churn. The horizontal axis indicates whether a customer has a cable connection while the vertical axis denotes the payment responsible (taking a "No" value when someone else than the customer, e.g. a parent, pays the bill). It appears that most Orange Belgium customers do not have a cable connection and are responsible for the payment. The color of the spots indicates the churn rate, with a lighter color denoting a higher probability of churn[2]. The impact of both binary

[1] *SIM-only* indicates that the customer bought no other product than the SIM card.

[2] For confidentiality reasons, the precise value of the churn rate cannot be disclosed.

variables appears clearly, with a significant difference in churn rate between the two extremes. The univariate impact of each variable on churn can be quantified in terms of odds ratio, measuring the increase of the odds of churn once exposed (i.e. when the customer is responsible for the payment, or when there is a cable connection). The odd ratios for the payment responsible and the cable connection are 0.917 and 0.839, respectively. This indicates that a “Yes” value for both variables is associated to a reduced risk of churn.

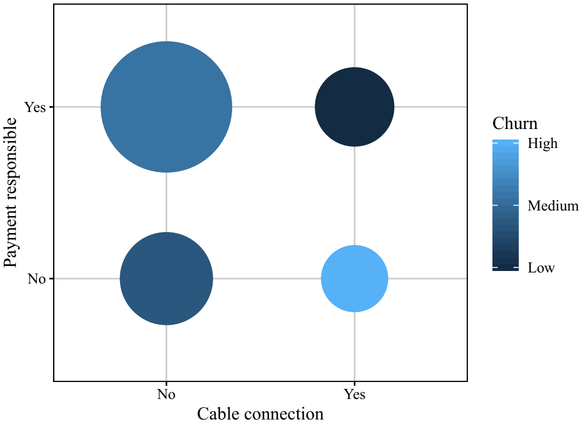


Fig. 1. Interaction between cable connection and payment responsible. A customer is not responsible for payment if someone else (e.g. a parent) pays the invoice in her stead. The color of the spots denotes the churn rate, whereas its area denotes the number of customers.

Another interesting visualization concerns the relation between tenure (i.e. the duration of the current subscription) and churn rate (Fig. 2)[3]. The curve shows a negative correlation between the churn rate and tenure. Note that the surge in churn rate corresponds to the term of the contract for loyalty customers.

2.2 Machine Learning Pipeline

Three different learning tasks are created by stratifying the dataset: one containing the loyalty contracts, one containing the SIM-only contracts, and one containing the SIM-only contracts with additional variables (denoted SIM-only Δ). The large unbalancedness of the dataset has been addressed by adopting the EasyEnsemble strategy [16] which consists in training a number (in our case 10) of learners on the whole set of positive instances (churners) and an equally sized random set of negative instances. Based on our previous experience on related

[3] For confidentiality reasons, the axes scales are concealed.

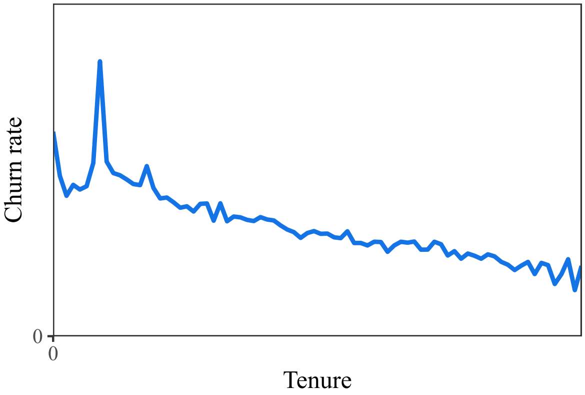


Fig. 2. Churn rate as a function of the tenure (the duration of the current subscription). The spike on the left of the plot corresponds to the end of the loyalty period for loyalty customers.

largely unbalanced tasks (notably fraud detection [3,4]) we considered as learner only Random Forests.

In what follows we report the results of a number of assessments evaluating the impact of

1. variable selection, based on the feature importance returned by Random Forest;
2. the addition of engineered features: for each time-dependent quantity (e.g. total duration of calls, or mobile data usage) we created 2 additional features measuring the difference and the ratio between two consequent monthly values, respectively;
3. the type of contract (SIM-only vs. loyalty).

The high computational cost of training on such a large dataset restricts the number of configurations we can assess. We limit the number of selected variables to 20, 30 or all variables. Also, we do not explore the difference variables for loyalty contracts. Overall we consider 9 different experiment configurations.

Three-fold cross-validation is used to assess the accuracy on the training set (first 4 months). The last month of data is used as a test set for each of the three datasets, in order to check the robustness of the prediction model (e.g. with respect to potential drifts or non-stationarity).

The performance of the different models is evaluated using three different measures: the receiver operating characteristic (ROC) curve, the precision-recall (PR) curve, and the lift curve [27]. While the ROC curve and the PR curve are widely used in conventional classification, the lift curve is of more practical interest in evaluating churn prediction. Since a customer churn retention campaign focuses on a limited amount of customers, the lift curve reports the expected performance of the model as the number of customers included in the campaign varies. From these 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% of customers included). There is some evidence in the literature [6,9,22] that the ROC curve is not a reliable metric on unbalanced data. Moreover, since the area under the ROC curve depends on all possible decision thresholds, it does not correspond with the objective of the churn campaign: finding a small group of customers with high churn probability (low false-positive rate). We report the AUROC to be consistent with the churn prediction literature, but our conclusions are mainly based on the other performance metrics.

2.3 Results and Discussion

Table 1 and 2 report the cross-validation and the test accuracy, respectively. Based on those results, a number of considerations can be made

- by reducing the number of features to 30, the accuracy does not deteriorate significantly. This is good news for our industrial partner since a compact churn model is more suitable for production.
- though adding engineered features may be beneficial, this occurs only if a feature selection is conducted beforehand.
- surprisingly, the accuracy is higher for the test set (Table 2) than in cross-validation (Table 1). Our interpretation, confirmed by visualization in the space of the two first principal components, is that the drift of the data makes the classification easier.
- regarding the type of contracts, churn is slightly easier to predict in the loyalty dataset than SIM-only, due to the greater importance of time-related variables. Indeed, the churn is significantly higher at the end of the mandatory period of a loyalty contract, facilitating the prediction process.

We compared our results on the SIM-only dataset with other published studies on churn prediction [5,12,18,28–31]. We achieve similar results in terms of area under the ROC curve and lift.

Table 1. Summary of the cross-validation results. Highest values for each type of contract and for each evaluation measure are underlined.

	SIM-only			SIM-onlyΔ			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 2. 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. Using only 20 variables decreases the performances most often.

	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

3 Causal Analysis

The selection procedure discussed in the previous section returns the most relevant variables for predicting the potential churners. Though such variables provide useful information for designing a predictor, they are not necessarily the ones to manipulate (e.g. by giving incentive) if we wish to reduce the churn risk. For example, an increase in the number of contracts registered by a customer may be strongly associated with a decrease of churn. However, a hypothetical churn retention action that would sell additional contracts might fail, if customer satisfaction has a causal effect both on the number of purchased contracts and the propensity to churn. In this case, the predictive variable (number of contracts) and the churn have a common latent cause (customer satisfaction). Manipulating the number of contracts will therefore not affect churn. Different tools are needed to discover true causal relationships between variables and will be discussed in what follows.

3.1 Causal Inference Strategy

We use the dataset of Sect. 2 and perform the causal inference separately on the SIM-only and loyalty customers 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. To decrease computation time, only the first 30 variables in the ranking of the random forest trained in Sect. 2 are used. A random undersampling has been applied to reach decent computation times and to perform class balancing. The positive class (churners) is kept fixed and a random subset of the negative class is randomly selected, such that the resulting dataset contains the same number of positive and negative observations. Further random undersampling is then performed, with a sample size depending on the inference algorithm.

The rationale of this experiment consists in applying several causal inference techniques, which give different types of results in various forms, and extract a consensus, if any, in the light of the different assumptions each model puts on the data. This strategy is called triangulating [14] and takes advantage of the fact that the different causal inference methods rely on different assumptions, thus increasing the validity of our results.

The considered causal inference algorithms are: PC [24], Grow-shrink (GS) [17], Incremental Association Markov Blanket (IAMB) [26], minimum Interaction Maximum Relevance (mIMR) [2] and D2C [1]. PC infers the equivalence class of causal graphs faithful to the dataset, GS and IAMB infer the Markov blanket of the churn variable, and mIMR and D2C return the direct causes of churn.

The PC [24] algorithm is slow when the number of samples is large since the whole causal network is inferred. Therefore, we restrict the dataset to 10,000 samples for this algorithm. The result is given under the form of a completed partial DAG (CPDAG) representing an equivalence class of directed acyclic graphs (DAG) [25].

The GS and IAMB algorithms [17,26] both infer the Markov blanket of a target variable, the churn in our case. These algorithms therefore return the direct causes, the direct effects and the direct causes of the direct effects (also called *spouses*). For these two algorithms, the entire set of positive samples is used, along with a subset of the same size of negative samples. IAMB differs from GS in that it is more sample-efficient.

Two implementations of the mIMR algorithm [2] are used: one based on histograms to estimate mutual information, and another assuming Gaussian distributions, thus allowing a closed-form formula for the computation of the mutual information [19]. For the first implementation, the dataset is restricted to 10,000 samples, due to the computational cost of the histogram-based estimator. In the second implementation, 100,000 samples from SIM-only are used, and all samples from loyalty are used. The results are provided as a list of the first 15 selected variables, accompanied by the gain provided by each variable at each iteration of the algorithm.

The D2C learning algorithm is trained using randomly generated DAGs, as described in [1]. 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 to be a cause of churn.

For the first three methods (PC, IAMB, and GS), we use the R package *bnlearn* [23] for independence tests using mutual information and asymptotic χ^2 test [8]. For mIMR and D2C, we use the R package *D2C* [1]. In all cases, a false-positive rate of 0.05 is chosen for statistical tests of independence.

Before proceeding with the results, it is worth reminding that all the 5 causal inference algorithms rely on specific assumptions. While PC, GS, and IAMB assume causal sufficiency and faithfulness, mIMR and D2C rely on more specific conditions.

Causal sufficiency denotes the absence of unmeasured confounder, and is likely to hold given the large number of variables (73) and the variety of information they provide (service usage, socio-demographic, type of subscription, etc). Confounding could be further reduced by including an indicator of service quality, which is absent from our dataset. See Sect. 3.3 for a more detailed review of our prior knowledge on the causes of churn.

The assumption of faithfulness states that any (conditional) independence found in the probability distribution is reflected by the d-separation of the relevant variables in the corresponding causal graph. Faithfulness in the case of the PC algorithm is discussed in [15].

mIMR is based on the assumption that direct causes form "unshielded collider" configurations together with the target. Since in such configurations direct causes are marginally independent and conditionally dependent, mIMR may exploit this pattern to prioritize direct classes in the ranking. Though such an assumption is hardly satisfied in real settings, the approach allows to introduce a causal criterion in a feature selection algorithm for large dimensional settings. The adoption of mIMR requires as well the choice of a mutual information estimator (typically Gaussian for its low-variance and robustness in non-normal configurations [19]).

D2C relies on the existence of asymmetric descriptors of the statistic dependency between a cause/effect pair. This is possible under some specific conditions, like the existence of a single edge connecting the Markov blanket of the cause and the effect and the existence of an unshielded collider between the cause (effect) and the related spouse. While the first assumption is probably not true in our setting, the second is satisfied by the fact that no descendant of the target variable is included in the dataset.

3.2 Sensitivity Analysis

Once causally relevant variables are inferred, it is worth evaluating the sensitivity of the target to their manipulation. In Sect. 2 we learned a predictive algorithm (random forest) to estimate $P(Y \mid X)$, i.e. the conditional probability distribution of the churn variable Y given the set of customer variables $\{X_1, \ldots, X_n\} = X$. Let us now focus on a putative cause $X_i \in X$ and assume causal sufficiency, i.e. all possible confounders are part of the set of measured variables X. In order to assess the sensitivity of Y to X_i, we measure the change in $P(Y \mid X)$ once the distribution of X_i is modified. This boils down to estimate the *natural direct effect* [21], which quantifies the causal effect of X_i on Y not mediated by any other variable, while the other variables are still distributed according to their natural distribution. This corresponds to answering the causal question "*What happens if only X_i changes?*".

Since we are interested in the effect of a shift in the distribution of X_i, we add $\alpha\sigma_i$ to the value of X_i, where σ_i is the standard deviation of X_i and α is a parameter of the intervention. The natural direct effect is

$$\text{NDE}_{x_i} = \text{NDE}_{x_i}(Y) - E[Y]$$

where $\mathrm{NDE}_{x_i}(Y)$ is defined as the expected value of Y under "natural" intervention on X_i, i.e. by letting other variables be distributed according to their original distribution:

$$\mathrm{NDE}_{x_i}(Y) = \int P(x)P(Y = 1 \mid x_1, \dots, \mathrm{do}(x_i + \alpha\sigma_i), \dots, x_n)\, \mathrm{d}x.$$

We know that all possible back-door paths between X_i and Y are blocked since, by causal sufficiency, $X = \{X_1, \dots, X_n\}$ includes all possible confounders. Therefore, using rule 2 of do-calculus [20] we can estimate $\mathrm{NDE}_{x_i}(Y)$ from observational data alone.

Given a dataset of n variables and N examples $\{(x_1^{(j)}, \dots, x_n^{(j)}; y^{(j)})\}_{1 \leq j \leq N}$, the average prediction of a model f on this dataset is an estimator of the expected value of Y:

$$E[Y] \approx \frac{1}{N} \sum_{j=1}^{N} f(x_1^{(j)}, \dots, x_n^{(j)})$$

For each variable X_i the expected value of Y under intervention can be approximated as

$$\mathrm{NDE}_{x_i}(Y) \approx \frac{1}{N} \sum_{j=1}^{N} f(x_1^{(j)}, \dots, x_i^{(j)} + \alpha\sigma_i, \dots, x_n^{(j)})$$

We applied this method on the SIM-only dataset, on the 30 variables having the largest importance according to the random forest models. The causal effect is computed for $\alpha = 1$ and $\alpha = -1$. The assumption of causal sufficiency (Sect. 3.1) is a necessary condition for the validity of this method. Note that the dataset we use in practice also contains discrete variables. These variables are left out of this analysis since the method is suited only to continuous variables.

3.3 Prior Knowledge

Before presenting the results of causal inference, it is interesting to summarize the knowledge of the Orange experts on the possible reasons for customer churn, elicited by means of several discussions and interviews. Those experts report four main causes of churn:

Bill shock: this occurs when a customer has an unusually large service usage, which results in an important "out of bundle" amount (i.e. the client is charged much more than usual). This triggers a reaction from the customer inducing an increased risk of churn. This scenario is well understood and verified in practice. It is believed to be the most important cause of churn.

Customer dissatisfaction: multiple factors influence customer satisfaction, including quality of service and network quality. A customer having numerous cuts of network connection during phone calls, or unable to use properly Orange online services, will be more likely to seek better alternatives elsewhere.

Wrong positioning: choosing the right tariff plan suited to one's service usage habits is sometimes difficult. On the one hand, if not enough call time is provisioned, an "out of bundle" amount is likely to be charged at the end of the month. On the other hand, an expensive tariff plan results in a high fixed cost for the customer. When the needs of a customer do not correspond to the chosen tariff plan, we say that the customer is wrongly positioned. A wrong positioning results in most cases to a higher bill than expected, and is a significant cause of churn.

Churn due to a move: it is common to choose a product bundle from a telecommunication company comprising a subscription for mobile phone, landline phone, television, and internet connection. In this case, the subscription is tied to the particular place of domicile of the customer. When the client moves to another place, it is quite common to also change for another telecommunication service provider. Therefore, this is a significant cause of churn, albeit of a different nature from the other settings exposed above.

While these potential causes of churn pertain to the whole customer base, the loyalty customers typically have a much higher churn rate at the end of the mandatory period of their contract, thus the tenure (the time since when a customer uses Orange' services) is an important cause of churn for them. On SIM-only customers, expert knowledge also indicates that the tenure influences churn: a new customer is more likely to churn than a long-time customer since it is less committed to the company.

These different settings are described informally, and their translation to the formal definitions of causality is not straightforward. We wish to find a mapping between the events believed to be causes of churn and specific instantiations of measurable random variables. In the case of the first setting (bill shock), we can reasonably assume that variables measuring the "out of bundle" amount of the customer is a faithful proxy for bill shock. Similarly, customer satisfaction can be estimated using, for example, the number of network cuts during phone calls, or the number of calls to the customer service. The wrong positioning can also be numerically estimated, given the tariff plan of the client and its average service usage. The last setting (churn due to a move) is much more difficult to account for, as it is not directly related to the interaction between the client and the telecommunication services.

In the dataset available for this study, the only measured variables that translate to potential causes of churn are the "out of bundle", the tariff plan and service usage (phone calls, messages, mobile data). We have no measure for network quality, customer satisfaction, or propensity to move soon. Also, the wrong positioning is not explicitly encoded and has to be inferred by the causal inference model from the average service usage and the current tariff plan.

3.4 Results of Causal Inference

The outcome of the inference algorithms is summarized in Figs. 3 and 4. Each of the possible causes of churn is represented by an ellipse, annotated with the algorithms that output this variable. For both SIM-only and loyalty, the PC algorithm infers an intricate causal graph but where the churn variable is disconnected from all others. Note that GS and IAMB output the Markov blanket, and not only direct causes. Since the output of mIMR is a ranking, we use background knowledge to determine how many of the top-ranked variables should be considered as inferred causes, based on their redundancy. In the case of the histogram-based mIMR, the first variables in the ranking are complementary, but the 10th variable (for SIM-only) and the 12th variable (for loyalty) are mostly redundant with the other variables higher in the ranking. This indicates that the variable interaction is low and the remaining variables lower in the ranking should not be considered as causes. For the mIMR with Gaussian assumption, there is a significant drop in the gain between the 7th and the 8th ranked variables in the SIM-only dataset, and between the 8th and 9th ranked variables in the loyalty dataset. We consider that as a criterion for considering only the 7 (SIM-only) and 8 (loyalty) first ranked variables as inferred causes. D2C outputs a probability of being a cause of churn, for each variable. For the SIM-only dataset, we selected the tariff plan, the province of residence and the data usage as causes inferred by D2C, and for the loyalty dataset, we selected the number of contracts, the province, and the tenure. These variables display a significantly higher predicted score than the other variables.

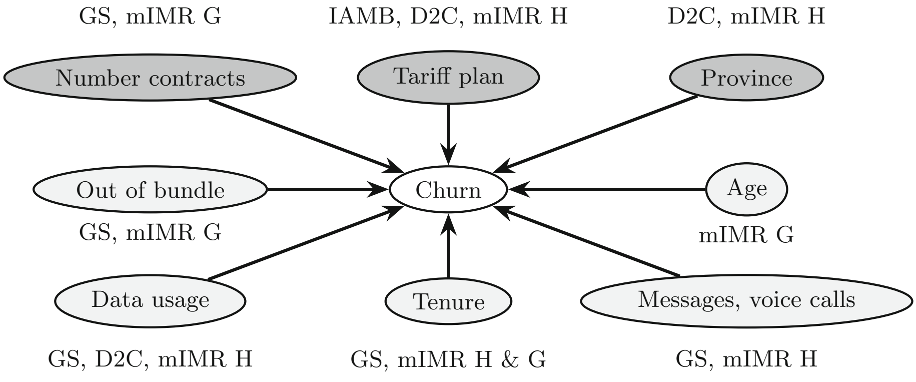


Fig. 3. Summary of results of causal inference on SIM-only dataset. Each variable is annotated with the algorithms predicting it to be a cause of churn. Light gray ellipses represent continuous variables, and dark gray ellipses represent discrete variables. mIMR H stands for the histogram-based estimator, and mIMR G for the estimator with Gaussian assumption.

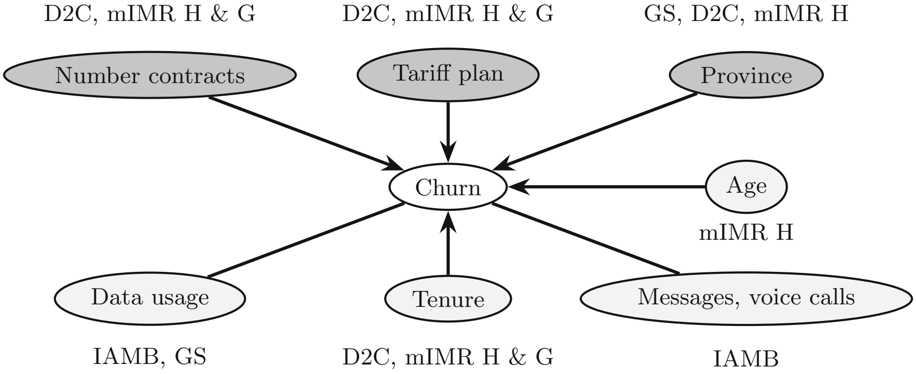


Fig. 4. Summary of results of causal inference on loyalty dataset. Each variable is annotated with the algorithms predicting it to be a cause of churn. Light gray ellipses represent continuous variables, and dark gray ellipses represent discrete variables. mIMR H stands for the histogram-based estimator, and mIMR G for the estimator with Gaussian assumption.

For the SIM-only dataset (Fig. 3), the “out of bundle” and data usage variables are reported as causes by mIMR and D2C, and as members of the Markov blanket by GS. This is in line with our prior belief that bill shock is a major cause of churn. We could expect the “out of bundle” variable to stand out more explicitly, but it is only given by mIMR with Gaussian assumption. However, the distribution of the “out of bundle” can roughly be modeled as the exponential of a Gaussian. It is thus easy to understand why the other inference methods, that make different statistical assumptions, fail to report the causal link to churn.

The tariff plan and the “out-of-bundle” variables together provide a representation of the tariff plan positioning of the customer. For the SIM-only dataset, these two variables are reported as causes of churn by mIMR and D2C and are also members of the Markov blanket according to GS and IAMB. This confirms our hypothesis that wrong positioning is an important cause of churn.

Note that the “out of bundle” is not reported by any algorithm for the loyalty dataset (Fig. 4). This is consistent with the fact that loyalty customers are not able to churn in the mandatory period of their contract, thus churn related to bill shock is less represented in this dataset.

The two last causes of churn according to Sect. 3.3 are customer satisfaction and churn due to a move. None of the measured variables are direct proxies for these two putative explanations of churn. Better results could be obtained by using relevant variables such as, for example, the number of calls to the customer service, a measure of the network quality, the number of network cuts during a call, and so on. Adding these variables would reduce latent confounding if the underlying causal hypotheses are true. We suspect that the importance of the province in Figs. 3 and 4 is an indication that network quality is an important cause of churn (the network quality is known to vary between different regions of Belgium). However, the scope of this study limited us to the set of variables presented in Sect. 2.1.

If we use the expert knowledge to assess the accuracy of the causal inference algorithms, mIMR H and D2C algorithms seem to better infer relevant variables as direct causes. Indeed, the bill shock and the wrong positioning imply that the "out of bundle", the tariff plan and the data usage are likely causes of churn. The two latter are output by mIMR H and D2C in the SIM-only dataset, whereas mIMR G outputs the "out of bundle". In the loyalty dataset, D2C and mIMR correctly avoid reporting the "out of bundle" or the data usage as causes of churn, but correctly report the importance of the tenure. A model similar to mIMR H or D2C, but able to correctly handle variables with more difficult distributions such as the "out of bundle" variable, would be ideal.

Finally, it is important to consider that these results may suffer from sampling bias. Given that we use a crude random undersampling technique, some causal patterns in the discarded positive samples may be under-represented in the resulting training set. This is especially the case for the PC algorithm (using 10,000 samples), the first implementation of mIMR (10,000 samples), and D2C (2,000 samples). And even though the remaining algorithms use far more samples, none of them can take into account the entire set of non-churners. Furthermore, we have no theoretical guarantee that an even class ratio is best for inferring causal patterns. Reducing sampling bias in causal analysis requires the conception of new techniques that are outside the scope of this article.

3.5 Results of Sensitivity Analysis

The results of the variable sensitivity analysis are shown in Figs. 5 and 6. Each variable is represented as a bar whose color depends on the category of variable: subscription, calls and messages, mobile data usage, revenue, customer hardware, and socio-demographic. Some variable names have been anonymized for confidentiality reasons. Also, only variables inducing the most significant change in the distribution are shown ($p < 10^{-10}$ with a two-sided t-test).

All the numerical variables inferred as possible causes of churn appear to influence the predictions of the model, albeit in a non-linear manner as indicated by the lack of symmetry between Figs. 5 and 6. On the one hand, the tenure and the number of contracts are observed to be monotonically associated with the churn probability, since they appear in both figures in opposite directions. On the other hand, variables related to the amount paid by the customer and the data usage cause more churn when they are increased, but the opposite is not true. Note that the tariff plan and the province, although reported as possible causes in Fig. 3, are not present in Figs. 5 and 6 since they are categorical, thus unsuitable for this analysis.

In Appendix A (Figs. 7 and 8), we report the entire distribution of predicted probability of churn for each shifted variable, instead of reporting only the difference between the means as in Fig. 6 and 5. This shows other characteristics of the causal impact of these variables on churn, such as the change in the spread of the probability distribution. Note that, while the churn is highly unbalanced in the original dataset, the predicted probability of churn is balanced. This is due to the EasyEnsemble methodology, which generates balanced subsets of the original training set.

The causal impact of a smaller intervention is reported in Appendix A, Figs. 9 and 10. The intervention consists in adding or subtracting $0.5\,\sigma_i$ from each variable separately, instead of σ_i as in Figs. 5, 6, 7, and 8. We observe that some variables have almost the same impact as with a shift of 1 sigma (e.g. the out of bundle variables), while others have significantly less impact, such as the number of contracts. In the latter case, this is due to the discrete distribution of the number of contracts. Other variables, such as D13, have a proportionally reduced impact on the predicted probability of churn.

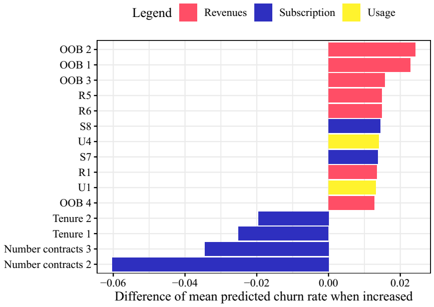


Fig. 5. Difference of mean predicted probability of churn when a standard deviation is added separately to each variable. Run on the SIM-only dataset. Only variables inducing the most significant change in the distribution are shown ($p < 10^{-10}$ with a two-sided t-test).

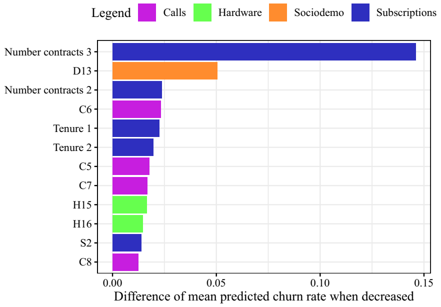


Fig. 6. Difference of mean predicted probability of churn when a standard deviation is subtracted separately from each variable. Run on the SIM-only dataset. Only variables inducing the most significant change in the distribution are shown ($p < 10^{-10}$ with a two-sided t-test).

4 Conclusion

Churn prediction in the telecommunication industry is notoriously a hard task characterized by the non-linearity of variables, large overlap between churners and non-churners, and class imbalance. Predictive modeling of churn was achieved with a random forest classifier and the Easy Ensemble algorithm. In a series of experiments on churn prediction, we assessed the impact of variable selection, type of contract and use of engineered features. The results show that variable selection helps reducing computation time if at least 30 features are selected. Also, the engineering of new features may be beneficial if variable selection is applied. We explored the application of causal inference from observational data. More specifically, we applied 5 different causal inference methods, namely PC, Grow-Shrink (GS), Incremental Association Markov Blanket (IAMB), minimum Interaction Maximum Relevance (mRMR), and D2C. The results of these algorithms are heterogeneous yet consistent with prior knowledge of the causes of churn. The direction of the causal influence of variables on churn was estimated through a novel method of sensitivity analysis. This method is based on the assumption that no latent variables are confounding factors of churn and the variable under inspection. This method showed that some variables have a non-monotonic causal influence on churn, which is consistent with expert knowledge.

5 Future Work

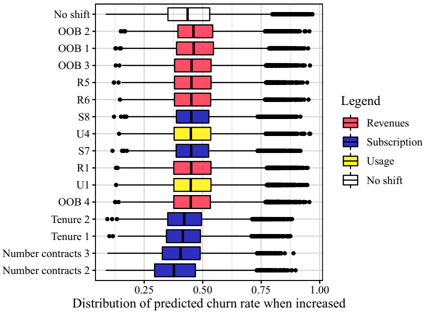
Results of causal analyses are difficult to validate without the ability to perform experiments. In this study, we are limited to compare our findings with prior knowledge of experts. Retention campaigns provide a promising opportunity to validate causal hypotheses. They can emulate a variable manipulation by offering risky customers targeted promotions. We plan to conduct such experiments in the future through collaboration with Orange Belgium.

Uplift modeling is an interesting approach to incorporate causal consideration in churn prediction [10]. In uplift modeling, the customers are ranked according to the diminution of their probability of churn when subject to the campaign, as opposed to usual churn modeling that ranks customers according to their probability of churn. Retention campaigns will allow assessing the effectiveness of this approach.

Another limitation of our approach is the arbitrary decision threshold we fixed between inferred causes and non-causes for the mIMR and D2C algorithms. Since these two methods output a score for each variable, we can instead compute a performance curve (*e.g.* ROC, precision-recall) from the predicted scores and the ground truth provided by experts. Although this is not suitable for performing causal discovery *per se*, this allows to quantitatively evaluate causal inference algorithms.

Undersampling and class balancing are used to ensure the computational tractability of causal inference. However, this may result in sampling bias, and its effect on the results has not been formally assessed. We can obtain more robust and stable results by performing undersampling and the causal inference experiments multiple times, as in the EasyEnsemble methodology.

A Additional Figures on Sensitivity Analysis



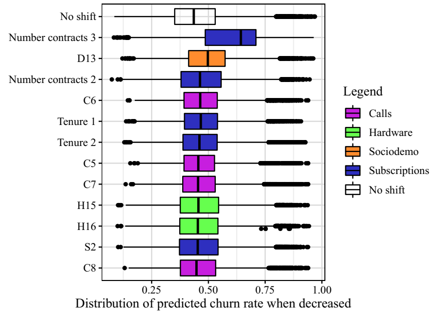
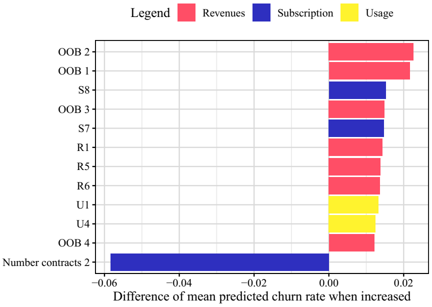


Fig. 7. Distribution of the predicted probability of churn when a standard deviation is added separately to each variable. Run on the SIM-only dataset. Only variables inducing the most significant change in the distribution are shown ($p < 10^{-10}$ with a two-sided t-test).

Fig. 8. Distribution of the predicted probability of churn when a standard deviation is subtracted separately from each variable. Run on the SIM-only dataset. Only variables inducing the most significant change in the distribution are shown ($p < 10^{-10}$ with a two-sided t-test).



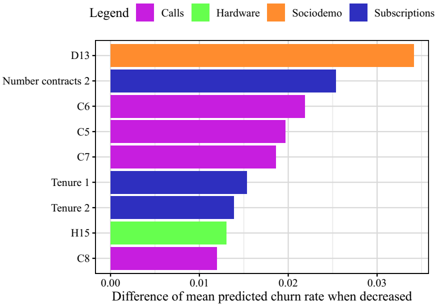


Fig. 9. Difference of mean predicted probability of churn when half a standard deviation is added separately to each variable. Run on the SIM-only dataset. Only variables inducing the most significant change in the distribution are shown ($p < 10^{-10}$ with a two-sided t-test).

Fig. 10. Difference of mean predicted probability of churn when half a standard deviation is subtracted separately from each variable. Run on the SIM-only dataset. Only variables inducing the most significant change in the distribution are shown ($p < 10^{-10}$ with a two-sided t-test).

References

1. Bontempi, G., Flauder, M.: From dependency to causality: a machine learning approach. J. Mach. Learn. Res. **16**(1), 2437–2457 (2015)
2. Bontempi, G., Meyer, P.E.: Causal filter selection in microarray data. In: Proceedings of the 27th International Conference on Machine Learning (icml-10), pp. 95–102 (2010)
3. Dal Pozzolo, A., Bontempi, G.: Adaptive machine learning for credit card fraud detection (2015)
4. Dal Pozzolo, A., Caelen, O., Waterschoot, S., Bontempi, G.: Racing for unbalanced methods selection. In: Yin, H., et al. (eds.) IDEAL 2013. LNCS, vol. 8206, pp. 24–31. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41278-3_4
5. De Caigny, A., Coussement, K., De Bock, K.W.: A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. Eur. J. Oper. Res. **269**(2), 760–772 (2018). https://doi.org/10.1016/j.ejor.2018.02.009
6. Elazmeh, W., Japkowicz, N., Matwin, S.: Evaluating misclassifications in imbalanced data. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) ECML 2006. LNCS (LNAI), vol. 4212, pp. 126–137. Springer, Heidelberg (2006). https://doi.org/10.1007/11871842_16
7. Fisher, R.A.: The Design of Experiments. Oliver and Boyd, Edinburgh, London (1937)
8. Good, P.: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer, New York (2013). https://doi.org/10.1007/978-1-4757-2346-5
9. Gu, Q., Zhu, L., Cai, Z.: Evaluation measures of the classification performance of imbalanced data sets. In: Cai, Z., Li, Z., Kang, Z., Liu, Y. (eds.) ISICA 2009. CCIS, vol. 51, pp. 461–471. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04962-0_53
10. Gutierrez, P., Gérardy, J.Y.: Causal inference and uplift modelling: a review of the literature. In: International Conference on Predictive Applications and APIs, pp. 1–13 (2017)
11. Hadden, J., Tiwari, A., Roy, R., Ruta, D.: Computer assisted customer churn management: state-of-the-art and future trends. Comput. Oper. Res. **34**(10), 2902–2917 (2007)
12. Idris, A., Khan, A.: Ensemble based efficient churn prediction model for telecom. In: 2014 12th International Conference on Frontiers of Information Technology (FIT), pp. 238–244 (2014). https://doi.org/10.1109/fit.2014.52
13. ITU: ITU releases 2018 global and regional ICT estimates (2018). https://www.itu.int/en/ITU-D/Statistics/Pages/stat/
14. Krieger, N., Davey Smith, G.: The tale wagged by the dag: broadening the scope of causal inference and explanation for epidemiology. Int. J. Epidemiol. **45**(6), 1787–1808 (2016)
15. Lemeire, J., Meganck, S., Cartella, F., Liu, T.: Conservative independence-based causal structure learning in absence of adjacency faithfulness. Int. J. Approx. Reason. **53**(9), 1305–1325 (2012)
16. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. Part B Cybern. **39**(2), 539–550 (2009). https://doi.org/10.1109/tsmcb.2008.2007853

17. Margaritis, D., Thrun, S.: Bayesian network induction via local neighborhoods. In: Advances in Neural Information Processing Systems, pp. 505–511 (2000)
18. Mitrović, S., Baesens, B., Lemahieu, W., De Weerdt, J.: On the operational efficiency of different feature types for telco Churn prediction. Eur. J. Oper. Res. **267**(3), 1141–1155 (2018). https://doi.org/10.1016/j.ejor.2017.12.015
19. Olsen, C., Meyer, P.E., Bontempi, G.: On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. EURASIP J. Bioinform. Syst. Biol. **2009**(1), 308959 (2008)
20. Pearl, J.: Causality: models, reasoning, and inference. IIE Trans. **34**(6), 583–589 (2002)
21. Petersen, M.L., Sinisi, S.E., van der Laan, M.J.: Estimation of direct causal effects. In: Epidemiology, pp. 276–284 (2006)
22. Raeder, T., Forman, G., Chawla, N.V.: Learning from imbalanced data: evaluation matters. In: Holmes, D.E., Jain, L.C. (eds.) Data Mining: Foundations and Intelligent Paradigms, pp. 315–331. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-23166-7_12
23. Scutari, M.: Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:0908.3817 (2009)
24. Spirtes, P., Glymour, C.: An algorithm for fast recovery of sparse causal graphs. Soc. Sci. Comput. Rev. **9**(1), 62–72 (1991)
25. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, vol. 81. Springer, New York (1993). https://doi.org/10.1007/978-1-4612-2748-9
26. Tsamardinos, I., Aliferis, C.F., Statnikov, A.R., Statnikov, E.: Algorithms for large scale markov blanket discovery. In: FLAIRS Conference, vol. 2, pp. 376–380 (2003)
27. Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B.: New insights into churn prediction in the telecommunication sector: a profit driven data mining approach. Eur. J. Oper. Res. **218**(1), 211–229 (2012)
28. Verbeke, W., Martens, D., Baesens, B.: Social network analysis for customer churn prediction. Appl. Soft Comput. **14**, 431–446 (2014). https://doi.org/10.1016/j.asoc.2013.09.017
29. Zhu, B., Baesens, B., vanden Broucke, S.K., : An empirical comparison of techniques for the class imbalance problem in churn prediction. Inf. Sci. **408**, 84–99 (2017). https://doi.org/10.1016/j.ins.2017.04.015
30. Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., Vanthienen, J.: Social network analytics for churn prediction in telco: model building, evaluation and network architecture. Expert Syst. Appl. **85**, 204–220 (2017). https://doi.org/10.1016/j.eswa.2017.05.028
31. Óskarsdóttir, M., Van Calster, T., Baesens, B., Lemahieu, W., Vanthienen, J.: Time series for early churn detection: Using similarity based classification for dynamic networks. Expert Syst. Appl. **106**, 55–65 (2018). https://doi.org/10.1016/j.eswa.2018.04.003