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] Machine Learning Group, Computer Science Department,
Université Libre de Bruxelles, Brussels, Belgium
{tverhels,gbonte}@ulb.ac.be
[2] Data science team, Orange 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. Retention campaigns can be used to prevent customer churn, but their effectiveness 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 allows to infer causal relationships between variables purely 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, assuming the absence of confounding factors.

Keywords: Churn prediction · Machine Learning · Big data · Causal inference.

1 Introduction

In recent years, the number of mobile phone users increased substantially, reaching more than 3 billion users worldwide. The number of mobile phone service subscriptions is actually greater than the number of residents in several countries, including Belgium [10]. Telecommunication companies are evolving in a saturated market, where their customers are exposed to competitive offers from many other companies.

Hadden et al. [8] 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. It is beneficial for the telecommunication company to focus the retention campaigns only on risky customers, in the hope of preventing attrition that would otherwise occur if no actions were taken. Churn detection using machine learning and data mining is nowadays performed by most major telecommunication companies, and a part of the data mining literature is devoted to churn prediction [5, 23, 21, 13, 9, 22, 20].

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. This first part of this paper assesses several machine learning methods and strategies in a large set of real historical data about the churn behavior of Orange Belgium telecom clients. We show that a machine learning pipeline can successfully predict potential churners, supporting the design of targeted retention campaigns. Estimating the probability of churn of a customer is however not sufficient to define an effective campaign if we do not know which specific incentive to propose to the potential churner. 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* [6], 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.

- Evaluation of the predictive power of a state-of-the-art churn prediction model, and the impact of several variations of the model by using different features and different type of subscription contracts (Section 2).
- Application of causal strategies to inference putative causes of churn from observational data (Section 3).
- Assessment of the direction of the impact of putative causal variables on predictions (Section 3).

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

2 Churn prediction

This section describes the real 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 the type of subscription, the hardware, the mobile data usage (in MB), the number of calls/messages and some socio-demographic information. The dataset has 7.6 million entries (about 1.5 million entries per month). The target variable, 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 imbalanced, meaning that there are far more non-churners than churners.

Two kinds of subscriptions are present in this dataset: SIM-only[3] and loyalty. The first type refers to a subscription where the customer can quit at any time with no cost. This is not the case for the second contract type where the customer receives a large discount on the purchase of a mobile phone but agrees not to churn for a certain time (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 remaining discount amount. 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 of the experiments will nevertheless be conducted on both types of contract, in order to understand the differences in terms of churn behavior.

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

2.2 Machine learning pipeline

The large unbalancedness of the dataset needs to be addressed [4]. Class balancing is achieved by adopting the EasyEnsemble strategy [11] which consists in training a number (in our case 10) of learners on the whole set of positive instances (churners) and on an equally sized random set of negative instances. Based on our previous experience on related largely unbalanced tasks (notably fraud detection [3]) we considered as learner only Random Forests. We explored several alternative configurations in terms of features and learning tasks.

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.

Three learning tasks are considered 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 $\varDelta$).

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 (e.g. difference and ratio variables);
3. the type of contract (SIM-only vs. loyalty).

The high computational cost of the 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 [19]. 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 allows observing 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%).

2.3 Results and discussion

Table 1 and 2 report the cross-validation and the test accuracy, respectively. On the basis of 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 a 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, 23, 21, 13, 9, 22, 20]. We achieve similar results in terms of area under the ROC curve and lift.

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

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

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.

3 Causal analysis

The variable selection procedure discussed in the previous section returns which variables are relevant for predicting the clients most likely about to churn. Though this information is useful for designing a good predictor, it is not necessarily useful in the perspective of an intervention (e.g. incentive) 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 on 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 same dataset as in section 2, restricting 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. To decrease computation time, only the first 30 variables in the ranking of the random forest trained in section 2 are used. A random subsampling has been applied to reach decent computation times and to perform class balancing.

The overall scheme 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. 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 these state-of-the-art causal inference algorithms: PC [17], Grow-shrink (GS) [12], Incremental Association Markov Blanket (IAMB) [18], Minimum interaction maximum relevance (mIMR) [2] and D2C [1].

The PC [17] algorithm is slow when the number of samples is large since the whole network is inferred. Therefore, we restrict the dataset to 10,000 samples for this algorithm. The result is given under the form of a directed acyclic graph.

The GS and IAMB algorithms [12, 18] both infer the Markov blanket of a target variable, the churn in our case. However, it is left unspecified how the members of the inferred Markov blanket causally relate to the target variable. For these two algorithms, the entire set of positive samples is used, along with a subset of the same size of negative samples.

Two implementations of the mIMR algorithm [2] are used: one based on histograms to estimate mutual information, and another assuming Gaussian variables, thus allowing a closed-form formula for the computation of the mutual information [14]. 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 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* [16] for independence tests using mutual information and asymptotic χ^2 test [7]. 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.

3.2 Sensitivity analysis

Besides the inference of causally relevant variables, we also present a novel method that evaluates the sensitivity of the target to these relevant variables. Consider a predictive algorithm such as a random forest used in section 2. The goal of such an algorithm is to estimate the probability distribution of the churn variable Y given the set of customer variables $\{X_1, \dots, X_n\} = X$, that is, $P(Y|X)$. Let us consider a variable $X_i \in X$, and assume, for the sake of simplicity, that X_i is a cause of Y without any confounding factor. In order to assess the sensitivity of Y to X_i, we consider how the learning algorithm modifies its estimation of $P(Y|X)$ when the distribution of X_i is modified. Because of the absence of confounding, this is a correct estimation of the effect of a manipulation $\mathrm{do}(X_i)$ [15].

In order to quantify the influence of X_i on the distribution of Y, we compute the difference in the expected value of Y with and without the intervention $\mathrm{do}(X_i)$. We simulate $\mathrm{do}(X_i)$ by adding or subtracting a standard deviation from all instances of X_i in a test dataset. More practically, given a dataset of n numerical 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

$$M = \frac{1}{N} \sum_{j=1}^{N} f(x_1^{(j)}, \dots, x_n^{(j)})$$

For each variable $i \in \{1, \dots, n\}$ in this dataset, we compute the *shifted average prediction*

$$M_i^{\pm} = \frac{1}{N} \sum_{j=1}^{N} f(x_1^{(j)}, \dots, x_{i-1}^{(j)}, x_i^{(j)} \pm \sigma_i, x_{i+1}^{(j)}, \dots, x_n^{(j)})$$

where σ_i is the standard deviation of variable i. The differences $M - M_i^+$ and $M - M_i^-$ indicate the impact of a small modification of variable i on churn predictions. We applied this method on the 30 variables having the largest importance according to the random forest models. The assumption of no confounding may not be met for all these variables, but it has the benefit of simplifying the analysis. Moreover, we can disregard non-causal associations using the results of the causal inference experiments. 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.

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, 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.

Expert knowledge also indicates that the tenure (the time since when a customer uses Orange' services) has a significant influence on churn. A new customer is more likely to churn than a long-time customer since it is less committed to the company.

3.4 Results of causal inference

The outcome of the inference algorithms is summarized in Figure 1. Each of the possible causes of churn is represented by an ellipse, annotated with the algorithms that output this variable. 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 9 variables in the ranking are complementary, but the 10th variable is mostly redundant with the 9th one. This indicates that the variable interaction is low and the remaining variables 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. We consider that as a criterion for considering only the 7 first ranked variables as inferred causes. D2C outputs a probability of being a cause of churn, for each variable. We selected the tariff plan, the province of residence and the data usage as causes inferred by D2C since the remaining variables display a significantly lower predicted score.

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 the 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. These two variables are reported as causes of churn by mIMR and D2C and are also members of the

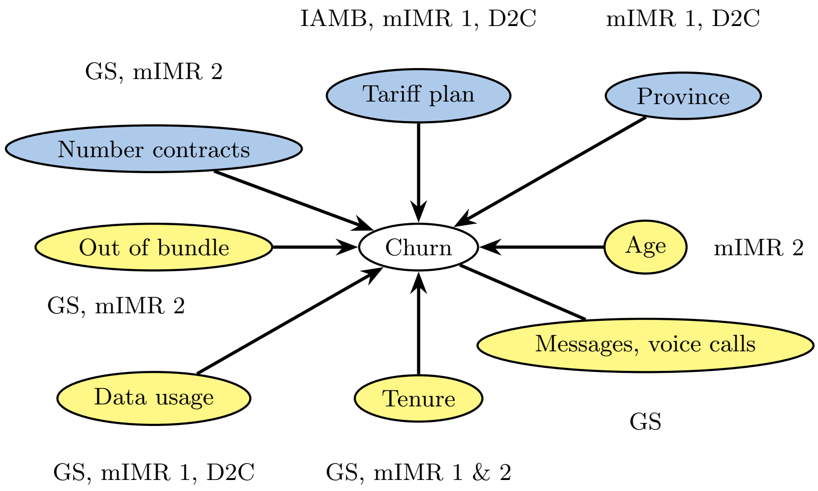

Fig. 1. Summary of results of causal inference. Each variable is annotated with the algorithms predicting it to be a cause of churn. Yellow ellipses represent continuous variables, and blue ellipses represent discrete variables. mIMR 1 stands for the histogram-based estimator, and mIMR 2 for the estimator with Gaussian assumption.

Markov blanket according to GS and IAMB. This confirms our hypothesis that wrong positioning is an important cause of churn.

The two last causes of churn according to section 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 variable 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. However, the scope of this study limited us to the set of variables presented in section 2.1.

If we use the expert knowledge to assess the accuracy of the causal inference algorithms, mIMR 1 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 1 and D2C, whereas mIMR 2 outputs the "out of bundle". A model similar to mIMR 1 or D2C, but able to correctly handle variables with an exponential distribution such as the "out of bundle", 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 figures 2 and 3. Each variable is represented as a bar whose color depends on the category of variable: subscription (yellow), calls and messages (blue), mobile data usage (green), revenue (purple), customer hardware (pink), and socio-demographic (orange). Some variables names have been anonymized for confidentiality reasons. Also, variables inducing a negligible shift in average predicted churn probability (respectively less than 0.01 and 0.005 for figures 2 and 3) are not reported.

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 figures 2 and 3. On the one hand, the tenure and the number of contracts (in yellow) 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 (in purple) and the data usage (green) 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 figure 1, are not present in figures 2 and 3 since they are categorical, thus unsuitable for the application of this algorithm.

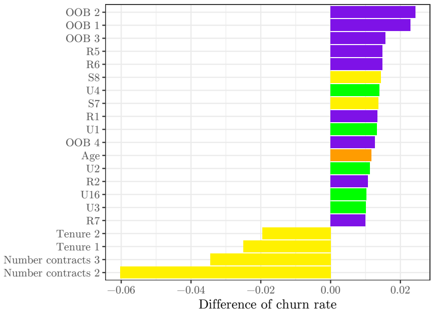

Fig. 2. Difference in the predicted probability of churn when a standard deviation is added separately to each variable. Run on the SIM only dataset. Only variables inducing a difference having an absolute value greater than 0.01 are shown.

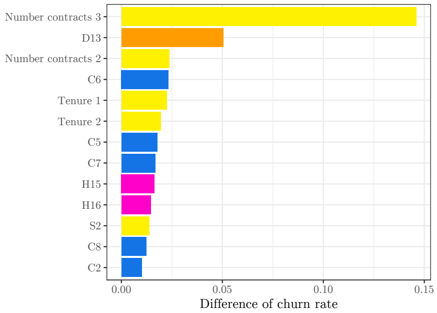

Fig. 3. Difference in the predicted probability of churn when a standard deviation is subtracted separately from each variable. Run on the SIM-only dataset. Only variables inducing a difference having an absolute value greater than 0.005 are shown.

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 varied and are consistent with prior knowledge of the causes of churn. The direction of the causal influence of variables on churn is estimated through a novel method of sensitivity analysis. This method is based on the assumption that no latent variables are confounding factor of churn and the variable under inspection. This method shows that some variables have a non-monotonic causal influence on churn, which is consistent with expert knowledge. Results of causal analysis 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 hypothesis. They can emulate a variable manipulation by offering risky customers targeted promotions. We plan to conduct such experiments in the future through a collaboration with the direct marketing department of Orange Belgium.

References

1. Bontempi, G., Flauder, M.: From dependency to causality: a machine learning approach. The Journal of Machine Learning Research **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: International Conference on Intelligent Data Engineering and Automated Learning. pp. 24–31. Springer (2013)
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. European Journal of Operational Research **269**(2), 760–772 (2018). https://doi.org/10.1016/j.ejor.2018.02.009
6. Fisher, R.A.: The design of experiments. Oliver And Boyd; Edinburgh; London (1937)

7. Good, P.: Permutation tests: a practical guide to resampling methods for testing hypotheses. Springer Science & Business Media (2013)
8. Hadden, J., Tiwari, A., Roy, R., Ruta, D.: Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research **34**(10), 2902–2917 (2007)
9. Idris, A., Khan, A.: Ensemble based efficient churn prediction model for telecom. In: Frontiers of Information Technology (FIT), 2014 12th International Conference on. pp. 238–244 (2014). https://doi.org/10.1109/fit.2014.52
10. ITU: Itu releases 2018 global and regional ict estimates (2018), https://www.itu.int/en/ITU-D/Statistics/Pages/stat/
11. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **39**(2), 539–550 (2009). https://doi.org/10.1109/tsmcb.2008.2007853
12. Margaritis, D., Thrun, S.: Bayesian network induction via local neighborhoods. In: Advances in neural information processing systems. pp. 505–511 (2000)
13. Mitrović, S., Baesens, B., Lemahieu, W., De Weerdt, J.: On the operational efficiency of different feature types for telco Churn prediction. European Journal of Operational Research **267**(3), 1141–1155 (2018). https://doi.org/10.1016/j.ejor.2017.12.015
14. Olsen, C., Meyer, P.E., Bontempi, G.: 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 (2008)
15. Pearl, J.: Causality: models, reasoning, and inference. IIE Transactions **34**(6), 583–589 (2002)
16. Scutari, M.: Learning bayesian networks with the bnlearn r package. arXiv preprint arXiv:0908.3817 (2009)
17. Spirtes, P., Glymour, C.: An algorithm for fast recovery of sparse causal graphs. Social science computer review **9**(1), 62–72 (1991)
18. 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)
19. 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. European Journal of Operational Research **218**(1), 211–229 (2012)
20. Verbeke, W., Martens, D., Baesens, B.: Social network analysis for customer churn prediction. Applied Soft Computing **14**, 431–446 (2014). https://doi.org/10.1016/j.asoc.2013.09.017
21. Zhu, B., Baesens, B., vanden Broucke, S.K.: An empirical comparison of techniques for the class imbalance problem in churn prediction. Information sciences **408**, 84–99 (2017). https://doi.org/10.1016/j.ins.2017.04.015
22. 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 Systems with Applications **85**, 204–220 (2017). https://doi.org/10.1016/j.eswa.2017.05.028
23. 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 Systems with Applications **106**, 55–65 (2018). https://doi.org/10.1016/j.eswa.2018.04.003