Thesis Abstract: Churn Prediction and Causal Analysis on Telecom Customer Data*

Théo Verhelst[1], Olivier Caelen[2],
Jean-Christophe Dewitte[2], and Gianluca Bontempi[1]

[1] Machine Learning Group, Computer Science Department,
Université Libre de Bruxelles, Brussels, Belgium
[2] Data science team, Orange Belgium

Telecommunication companies are evolving in a highly competitive market where customers are exposed to many competitive offers from other companies. Hadden et al. [4] showed that attracting new customers can be up to six times more expensive than retaining existing ones. Retention campaigns are usually used to reduce customer churn, but their effectiveness depends on the accuracy of churn prediction models. The problem of churn prediction is notably difficult, as it involves large amounts of data, non-linearity, imbalance and low separability between churners and non-churners.

In this master thesis, we approach the churn prediction problem with Orange Belgium customer data. The dataset is a monthly summary of Orange Belgium customers' activity covering a 5 months time window in 2018. A descriptive analysis of the dataset is conducted, highlighting a large variety of variable types and distributions. We also demonstrate interactions between categorical variables. The churn prediction problem is highly imbalanced, meaning that there are far more non-churners than churners.

The large class imbalance between the classes of churners and non-churners is addressed with the EasyEnsemble strategy [5] which consists in training a number (in our case 10) of learners on the whole set of positive instances (churners) and on an equal-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. 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. The importance of these engineered features is evaluated. We also assess the impact of feature selection on prediction performances.

We observe that feature selection can be used to reduce computation time and memory requirements without deteriorating the performance if at least 30 variables are selected. This is an important result for our industrial partner since a compact churn model is more suitable for production. The addition of engineered variables can improve performance if feature selection is conducted before the training phase.

We explore the application of data-driven causal inference, which allows inferring causal relationships between variables purely from observational data. The

different inference algorithms give results in various form, thereby preventing a direct comparison. Thus, we adopt a "wisdom of the crowd" approach by applying all algorithms in parallel and combining their results for final considerations. More specifically, we apply PC [7], Grow-Shrink (GS) [6], Incremental Association Markov Blanket (IAMB) [8], minimum Interaction Maximum Relevance (mRMR) [2], and D2C [1].

The results of the experiments conducted with these algorithms are varied, but a consensus can be extracted. The bill shock and the wrong tariff plan positioning are likely hypotheses of churn. The tenure (the total duration since when a customer uses Orange's services) has also an influence on churn. This is consistent with prior knowledge of experts at Orange Belgium on the causes of churn. We could obtain better results by using a dataset where putative causes are directly represented as variables, thus reducing latent confounding.

The direction of the impact of putative causes of churn is estimated through a sensitivity analysis. This represents an original contribution of this master thesis. We calculate the shift in the average predicted probability of churn when a constant value is added or subtracted to each putative cause. Assuming the absence of latent confounding, this is a correct estimation of an intervention on the cause. The results of this experiment highlight the non-linearity of the influence of relevant variables on churn. On the one hand, the tenure and the number of contracts are observed to be monotonically associated with the churn probability. On the other hand, variables related to the amount paid by the customer and the data usage cause more churn when they are increased and have no effect on churn when they are decreased.

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. 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)
5. 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
6. Margaritis, D., Thrun, S.: Bayesian network induction via local neighborhoods. In: Advances in neural information processing systems. pp. 505–511 (2000)
7. Spirtes, P., Glymour, C.: An algorithm for fast recovery of sparse causal graphs. Social science computer review **9**(1), 62–72 (1991)
8. 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)