University of Rennes 1 - ENSAI

Report Synthesis

Relevance of the prediction of a solicitation of the infrastructure operator prior to a unitary intervention for the general public

Student name :
Diane Maillot-Tchofo

Academic tutor :
Catherine Benjamin

Company tutor :
Alain Petillon

31 august 2022

1 Introduction

Orange is one of the leaders in the field of telecommunications in France and more globally in Europe. The diversity of Orange's customer base and activities requires the various players and units to work on a wide variety of subjects, more or less close to the end customer.

My internship took place within the team named **B**oucle **L**ocale, **E**xperience client, **D**iagnostic et **D**ata Mining (BLEDD). Composed of 16 people with very disparate academic formations and experiences. 3 people work only on the optical fiber network (optical local loop[1]). One persone only work on the copper network, also called ADSL (copper local loop, CLL). 2 people work on the production of data bases. Finally, the other member of the unit work on issues that mixes both types of network, mainly to improve the global operation of the local loop.

1.1 Problematic

From the data made available by the company, these 6 months objective was to define, in the ideal case, a predictive model. This model's aim is to predict a specific type of intervention closure on the fiber network. If this objective is not met; we aim to at least produce a data table, including the code that produced it, and the reasoning behind the table's construction.
The type of failure targeted here is relatively subtle, it is actually still subject to misinterpretation from people of the technical staff, or from external parties.
The fiber network, as well as the copper network, is composed of multiple physical infrastructures. We will cite the three that are the most interesting to us :

- the optical junction node (NRO) is the farthest equipment from the client,
- the branchement point (PB) which is the closest to the client, for example inside a building if it shelters more than 12 dwellings,
- the mutualisation point (PM) is located between the NRO and the PB.

In France, since the end of France Telecom's monopoly in the Networks and Telecommunications sector, the exploitation of these 3 components is ruled by the ARCEP[2]. Since then, there has been 2 types of operators : commercial operators (OC) and infrastructure operators. Some OI have both hats, we can name Orange or SFR, their two types of activities have to be legally separated. Technically it means that a technician working for an OC **cannot** manipulate, repair or even touch the OI's equipments. The same obligations are applied to a technician working for an OI. However, if a failure is located at the PB, only an OI technician is accredited to repair it. In contrast, the PM has 2 doors, the equipment behind the left door is under the responsibility of the OC and the right door under the responsibility of the OI.

In this context, when a failure occurs on an equipment under the responsibility of the OI, an OC technician must first move to physically locate the failure, then he has to notify it to the corresponding OI whom has, in turn, to call upon one of its technician, and so on. As one can guess, this process is long and expensive. Long for the client which find himself without any service during sometimes several months. Expensive for OCs that have to offer a back up solution (more often than not very expensive) to its clients, but also pay the technical staff. Furthermore that these interventions cannot be charged to the client.

This very special kind of failure accounts for around 10% of all interventions of the commercial operator Orange (OC Orange), at a rhythm of 75 euros per interventions, many of these failures needs at the minimum 2 interventions, meaning 150 euros for one client and one failure. Which, for a relatively small part of the traffic account for enormous spending.

The objective was to predict if a client's failure is or is not due to OI equipment(s). This task requires to take into consideration several angles. Not only for the modelling part, but also the data table creation and more importantly the examination of the pertinence of a model deployment. The latter and the first were the most scrutinised given the potential implications.

In this summary we will give a brief overview of what has been done during the internship and some of the conclusions we have come to.

[1] The optical local loop, OLL, is the part of the network that carries the flow of the data from the operator's first equipment to the client's home.

[2] Autorité de Régulation des Communications Électroniques, des Postes et de la Distribution de la Presse.

2 Data Table

Orange is one of the companies that host (at least part of) their data themselves. Therefore, to construct a data suitable for both analysis and modelling, a complete access to compliant data was given to me. The richness and volume of available data are such that it took me several days (to weeks) to fully be able to maneuver the few source tables necessary to mine. In the end, almost 3 months were necessary to construct the table.
Indeed, the latter has to respond to very specific needs. Given that no such table was available before, I had to start from "scratch", which means that multiple meetings with experts and data analysts/scientists had to be scheduled. In the end, it was clear that to be viable, the table would need to respect a few rules. They are given below.

- Be stable : its variables are preferably sourced from tables known for their qualities (reliable, regularly updated, etc..) and that are not or very rarely modified, so that a variable that we choose from them does not disappear after a few weeks.
- Be easily checked : the calculated fields must be check-able for whom knows a specific case, for example area or a given intervention.
- Be reproducible : with the help of provided scripts (3 in total are needed, in 2 different programming languages) and the full report of this internship one should be able to reproduce the table and modify it at will.
- Allow modelling and analysis : one of the end goal is to brings complete or partial answers to more general issues.
- Respect GDPR : it is only allowed to keep a 1-year data historic to comply with regulations.

In the end, the table we constructed contains more than 400 000 interventions and 123 variables, including 74 numerical, 29 categorical and 20 are time-based.

3 Modelling

As already mentioned what we try to predict accounts for 10% of all interventions on the fiber network. Thus, our modelling phase was hugely oriented by this small number. The type of failure we are interested in depend on the conclusion of the on-site technician. The latter has to answer some questions about the intervention just conducted. If he wants to declare an OI failure, then the intervention closure will be "TSO" for Transfer to Service Operator.

Our modelling can be seen as a little complicated. As a first step we opted for a non supervised classification. The underlying idea was to have, as an output, multiple classes and some containing, acceptable level of TSOs so that we could do intra-cluster binary classification afterwards. Given our data structure and their distributions, ending up with clusters with 35% of TSOs or above did not seem unattainable. In this clustering step, we expected to see emerged more or less the same distinctions inter-clusters that an expert would have done, but with potentially other thresholds.

The first step of our modelling phase is therefore made of a clustering algorithm. to be in line with our objective we used percentages of TSO in cluster as the primary choice metric. We then have 11 classes, made by a gaussian mixture model. The class with the highest percentage of TSO is at 42%, the second at 35% and third and fourth at respectively 29 & 28 %. Conversely, the classes with the most interventions have between 5 and 7% of TSO, which is well below the global average.

Our modelling training phase can be summed in these 4 following steps :

- Data preparation, discretisation and scaling of categorical and numerical variables.
- Clustering with a Gaussian mixture model.
- Data preparation : transformation of the variable "closure code" to turn it into a binary variable (1 if the intervention was closed TSO, 0 if other).
- Binary prediction solely intra-cluster with the most percentages of TSO.

Hence, the binary prediction step is only applied to a restricted number of interventions. Given the clustering model we chose and its specifications, we applied the prediction to only 2 clusters ; we will call them cluster A and B for convenience. Cluster A has the most TSO interventions with 42%, the B one has 35%.

For reference, cluster A is predominantly made of clients residing in buildings (70%) rather than houses, in the greater Paris area (Ile de France department, 76%). The clients are connected to old or even very old PM, and those are connected to a lot of clients. They also are in areas where a lot of their neighbours have had TSO closures in the past. In total, this clusters is composed of 2 495 interventions.

Lastly, the average performances of our best models for each clusters are as follows. A benchmark model in the form of a prediction on the entire training set post clustering was also computed as a way of measuring the gain of doing the prediction inside each cluster.

- *benchmark (extra trees), 10.8% TSO closures : Accuracy : 0.91, Recall : 0.33, F1-score : 0.44, MCC : 0.43, Precision : 0.67.*
- Cluster B (Random Forest), 35% TSO closures : Accuracy : 0.78, Recall : 0.53, F1-score : 0.62, MCC : 0.49, Precision : 0.76.
- Cluster A (Random Forest), 42% TSO closures : Accuracy : 0.77, Recall : 0.65, F1-score : 0.70, MCC : 0.52, Precision : 0.76

4 Pertinence of Deployment

The relevance of a model deployment in a company like Orange France largely depends on the conjuncture. Not only the relevance but our data are equally affected by the economic and social situation, and if our data are, then as is the whole modelling phase. Among those issues, we can cite the following.

- The unemployment rate, in France in 2022 has reached its lowest number since 2008. This means that some companies now struggle to hire, and if they do the recruits are not as well trained as previously, thus interventions closure forms are impacted, but also how the interventions are conducted. This harms our data table as a whole.
- Contractual conflicts between Orange France and its subcontractors has been shown to hamper the quality of the data.
- New areas of the country are being linked to the fiber network. These areas will bring up different kind of challenges from the ones we have now. Therefore, a switch will soon happen, which would make the work that has been done on this much less of a priority.

Given the instability of our data, the current focus of the direction of Orange France combined to an unappropriated economic and social conjuncture, we conclude that it is not relevant, at this stage, to go further with the research done during these 6 months. However, the foundation are set for someone to pick it up when its more favourable.

To end this summary, this internship allowed me to learn how to navigate in a team but also within a big company. The amounts of tools available, the abundance of data, and so on. I also learned a new database query programming language, as well as how to use a number of new software, including Tableau or DBVizualizer. It also allowed me to solidify my professional project, both in the long and short term.