Linking Radio Access Network QoE and QoS with Ensemble Multiple Regression

Adrien Schaffner
LivingObjects
Toulouse, France
adrien.schaffner@livingobjects.com

Louise Travé-Massuyès
LAAS-CNRS & ANITI, University of Toulouse
Toulouse, France
louise.trave@cnrs.fr

Simon Pachy
LivingObjects
Toulouse, France
simon.pachy@livingobjects.com

Bertrand Le Marec
LivingObjects
Toulouse, France
bertrand.lemarec@livingobjects.com

***Abstract*—The evaluation of user satisfaction is an essential performance indicator for network operators. It can be impacted by several causes, including causes linked to the network. However, linking the subjective comments of a customer with an objective behavior of the network is an issue. Experience shows that an indicator taken from customer complaints gives a good trend on the level of network quality perceived by customers, but it is difficult to transpose into concrete actions because it is often unrelated to the key performance indicators on which engineers base their action plans. The objective of this work is to learn a model that links the complaint rate, expressed by the *Customer Satisfaction Rate* indicator, with a set of key performance indicators so that performance engineers better understand customer expectations and act foremost on the indicators that give the most dissatisfaction. To this end, this paper takes advantage of ensemble learning applied to multiple regression, focusing the ensemble strategy on variable selection. The model hence makes it possible to link Quality of Experience and Quality of Service, which is demonstrated by nice interpretable results obtained from applying the method to data from a French telecom case study.**

Index Terms*—Ensemble learning; Regression models; Data analysis; Knowledge extraction; Radio access networks; QoS/QoE relationship; Quality via QoE and customer reports.

I. Introduction

In the space of a few years, the telecom market has undergone numerous technological and regulatory transformations that have engendered a price war from which operators are now trying to get out. They try to better differentiate themselves by moving towards a better customer experience and better support. The evaluation criteria most often adopted to establish a comparison of mobile networks are field measurement campaigns or user satisfaction surveys. User satisfaction surveys are expressed by the number of complaints received, the presence or absence of unfair terms in contracts, the commercial network and telephone assistance, connection time as well as call drop rate and their management noted by a supervisory authority, such as ARCEP (Regulatory Authority for Electronic Communications and Posts) in France or FCC (Federal Communications Commission) in United States.

The Customer Satisfaction Rate (CSR) is a good performance indicator that helps operators to effectively manage and control their business and decision making. The CSR provides the number of complaints relative to the number of customers for a given area. However, predicting customer behavior, their level of satisfaction (or dissatisfaction) has always been a challenge for operators. It is therefore important to link the CSR to a set of Key Performance Indicators (KPI) that can easily be interpreted by performance engineers to act on the relevant causes of dissatisfaction.

This paper, whose beginnings can be found in [1], presents how to learn a model that links the CSR to a set of KPIs from data while selecting a set of explanatory KPIs from an oversized, but yet relevant, set. Compared to [1], the problem is cast into an ensemble learning framework. Adopting an original point of view, model prediction and variable selection are optimized in an interlinked way by an ensemble multiple regression process. This process considers a set of base models whose results are then combined. Unlike standard approaches, ensemble integration is focused on combining the variable selection results issued from the base models rather than directly the predictions. The final regression model captures the relationship between Quality of Experience (QoE) and Quality of Service (QoS).

The contents of the paper are organized as follows. Section II analyzes related work and positions the method of this paper with respect to the state of the art. Section III formulates the problem as a regression problem and provides the identified issues. Section IV presents two regression methods, Ordinary Least Squares (OLS) and Least Absolute Shrinkage and Selection Operator (LASSO), that are later used in the three base methods for ensemble generation in Section VI. Section V describes the application that aims at explaining the customer complaint indicator CSR that has been driving the design of the method. It also presents the data that has been used and the KPIs that have been considered as candidate explanatory variables. Section VII explains the steps of the ensemble integration method. The results of applying the ensemble integration method to the CSR problem are then interpreted in Section VIII. Finally, Section IX concludes the paper.

II. Related work

Much research investigated about customer complaint behavior since long [2] [3]. The idea of using complaint data to solve problems in design, marketing, installation, distribution and after sale use and maintenance, is quite natural. Understanding of customer complaint and market behavior has also been investigated so as to provide a framework for interpreting the data and extrapolating it to an entire customer base [4]. Especially in the mobile telecom industry, studies on customer complaint behaviour are numerous and continue today, significantly accentuated by the emphasis on machine learning techniques.

Given the increased competitiveness in this field, many studies have focused on a problem related to customer complaints, which is customer churn. Due to the direct effect on the revenues of the companies, especially in the telecom field, companies are seeking to develop means to predict potential customer to churn. Over the years, many machine learning algorithms have been used to produce churn prediction models and building feature's engineering and selection methods [5] [6] [7]. In the churn problem, not only complaint data but Henley segmentation, call details, line information, bill and payment information, account information, demographic profiles, service orders, etc. are potentially important. In this huge set of features, [8] identifies a subset of relevant features and applies several prediction techniques including Logistic Regressions, Multilayer Perceptron Neural Networks, Support Vector Machines and the Evolutionary Data Mining Algorithm in customer churn as predictors, based on the subset of features. [9] uses classification like the Random Forest algorithm, as well as, clustering techniques to identify the churn customers and provide the factors behind the churning of customers by categorizing the churn customers in groups.

In this paper, the focus is put on using solely complaint data to solve problems in maintenance. To do so, this work aims at linking the complaint rate with a set of technical KPIs that point at the cause of the complaints and suggest reconfiguration or repair actions on the network. This problem is much less explored in the literature than that of the churn. Literature can be exemplified by [10] that achieves correlation analysis and prediction between mobile phone users complaints and telecom equipment failures in three steps involving hierarchical clustering, pattern mining, and decision trees. On the other hand, [11] uses four machine learning algorithms, Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Decision Tree (DT) experimented on a database of 10,000 Korean mobile market subscribers and the variables of gender, age, device manufacturer, service quality, and complaint status. It found that ANN's prediction performance outperformed other algorithms. This last work takes into account much more data than those fixed by the objective of this paper. In addition, the first focuses on equipment failure while we want to handle the KPIs that are the data used on a daily basis by network monitoring operators. Last but not least, the algorithms used in [11] are certainly good for prediction, but they are limited in their ability to explain predictions. The relation between the prediction and the inputs of the model remains implicit. On the contrary, the objective of this work is to clearly explain this link so that it provides useful information. This is why, the approach has been based on simple regression models while the complexity of the problem is tackled with ensemble learning.

Ensemble learning is an active research topic in different communities, including pattern recognition, machine learning, statistics and neural networks [12]. Ensemble learning [13] relies on combining several learning algorithms to obtain better predictive performance, in particular in terms of robustness and accuracy [14]. Most works on ensemble learning focus on classification problems, however this approach can as well be interesting for regression problems. It is for this latter purpose that we are concerned with it.

In this paper, we adopt the general definition of ensemble learning proposed in [15]:

Definition 1 (Ensemble learning): Ensemble learning is a process that uses a set of models, each of them obtained by applying a learning process to a given problem. This set of models (ensemble) is integrated in some way to obtain the final prediction.

The direct approach to ensemble learning is managed in two steps: ensemble generation that generates a set of models and ensemble integration that implements a strategy for combining the prediction results of the base models [16]. This paper adopts an original point of view in considering two tasks at once: prediction and variable selection. Unlike standard approaches, ensemble integration is focused on combining the variable selection results issued from the base models rather than directly the predictions.

III. Problem formulation

In our approach, the problem of explaining the level of customer satisfaction (or dissatisfaction), i.e., the QoE, is formulated as the one of obtaining a model linking the CSR to a set of KPIs that can be interpreted by performance engineers in terms of operational actions, i.e., improving QoS. To obtain this model, we rely on multiple linear regression theory and cope with the complexity of the problem through ensemble learning.

Multiple linear regression [17] is a classic family of learning algorithms that postulates that a variable is expressed as the weighted sum of other variables. Multiple linear regression defines the conditions and the model according to which a quantitative variable y is explained by several other quantitative variables $x_j, j = 1, \ldots, p$. y is considered *dependent* or *endogenous* and the variables $x_j, j = 1, \ldots, p$ are said to be *explanatory* or *predictor* variables. Multiple linear regression assumes that the variation of each explanatory variable has an influence, with not necessarily equal proportions, on the behavior of the dependent variable. The function that relates the dependent variable to the explanatory variables is linear.

Summarizing, multiple linear regression is a learning method that postulates that a variable y (in our problem y=CSR) is expressed as the weighted sum of other variables. In our problem, we want to learn the relationship between some KPIs and the CSR, so that performance engineers better identify the causes of customer dissatisfaction and act first and foremost on the indicators that most influence. Formally, for a number p of explanatory KPIs named $x_j, j = 1, \ldots, p$, which are instanciated in Section V, and the dependent variable $y = CSR$, the goal is to learn weights $\beta_0, \beta_1, ..., \beta_p$ such as:

$$y = \beta_0 + \beta_1 x_1 + ...\beta_p x_p \tag{1}$$

For this, we have a dataset gathering n observed samples, $n > p + 1$, each of dimension $(p + 1)$ and identified by the index i:

$$(x_1^i, x_2^i, \ldots, x_p^i, y^i),\ i = 1, \ldots, n. \tag{2}$$

Observed samples are used to estimate the parameters $\beta_k, k = 0, \ldots, p$, that are assumed to be constant. Each sample is assumed to satisfy relation (1) with an error ϵ_i:

$$y^i = \beta_0 + \beta_1 x_1^i + ...\beta_p x_p^i + \epsilon_i,\ i = 1, \ldots, n. \tag{3}$$

Under some statistical assumptions on the error terms ϵ_i, in particular independence and identical distribution, the vector of parameters $\beta = (\beta_1, \ldots, \beta_p)^T$ and the nuisance parameter σ^2 defining the variance of the error $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$, i.e., $var(\epsilon) = \sigma^2 I$, can be estimated by classical methods like least squares minimization [18] or, assuming that the error terms follow a centered normal distribution, likelihood maximization [19].

The model obtained after estimation of the parameters can be evaluated by the coefficient of determination R^2.

$$R^2 = \frac{SSR}{SST} = \frac{\sum_i^n (\hat{y}^i - \bar{y})^2}{\sum_i^n (y^i - \bar{y})^2} \tag{4}$$

where $\hat{y}^i$ is the prediction for the i-th sample, $\bar{y}$ is the mean, SSR is the sum of squares due to regression, i.e., the variability from the mean $\bar{y}$ that the regression manages to explain, and SST is the sum of squares total, i.e., the variability of the observed variables around the mean.

R^2 represents the proportion of variance for the dependent variable that is explained by explanatory variables in the regression model. The closer the value of R^2 is to 1, the better the regression. However, in practice, the threshold value for R^2 for considering a good regression is highly dependent on the problem.

In our problem, the goal of the ensemble integrated regression model is to extract knowledge, i.e., to determine the KPIs that influence the CSR and to use the coefficients of the regression to quantify their influence on the CSR.

In practice, the issues to be faced are the following :

- Business experience tells us that each of the explanatory KPIs can only worsen the condition of the telecom network and therefore should increase the CSR (e.g., an increase in the call drop rate, in the expert's mind, naturally increases the CSR). It is hence important to take care of the signs of the coefficients obtained by the regression.
- The number of candidate KPIs for explanation is high and can lead to irrelevant models.

The last issue defines one of the main objectives of this work. Indeed, there are two important elements in a model to highlight the relationship between explanatory KPIs and the dependent variable CSR:

1) Which are the relevant explanatory KPIs ?
2) How strong is their influence ?

These two elements will come as the result of the ensemble regression method that we propose in Sections VI and VII.

IV. Two classical linear regression methods

This section presents the principles of two classical multiple regression methods that are used to obtain base models as presented in Section VI. These are then leveraged in the proposed ensemble integration method presented in Section VII.

A. Ordinary Least Squares

When trained with data, the Ordinary Least Squares (OLS) method [20] selects parameter values $\beta_j,\ j = 1, \ldots, p$ of the linear expression (1) by the principle of *least squares*. It minimizes the sum of the squares of the differences between the observed dependent variable value in the observed data $y^i, i = 1, \ldots, n$, and the value predicted by the linear function of the explanatory variables $\hat{y}^i, i = 1, \ldots, n$. The optimization criterion, or loss function, is thus given by:

$$\begin{aligned}\mathcal{L} &= \min_{\beta_0, \beta_1, \ldots, \beta_p} \frac{1}{2} \sum_{i=1}^{n} (y^i - \hat{y}^i)^2 \\ &= \min_{\beta_0, \beta_1, \ldots, \beta_n} \frac{1}{2} \sum_{i=1}^{n} (y^i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j^i)^2\end{aligned} \tag{5}$$

In geometrical terms, this can be seen as the sum of the squared distances, parallel to the axis of the dependent variable, between each data point in the set and the corresponding point on the regression surface. The smaller the differences, the better the model fits the data.

The OLS estimator is consistent, i.e., has convergence to the real parameters values as the training data is increased, when the regressors are exogenous. It is optimal in the class of linear unbiased estimators when the errors are homoscedastic, i.e., they have the same variance, and are serially uncorrelated. Under these conditions, the OLS method provides minimum-variance mean-unbiased estimation when the errors have finite variances. Under the additional assumption that the errors are normally distributed, OLS is the maximum likelihood estimator.

In this work, the function `ols` of the Python module `statsmodels` has been used to implement OLS.

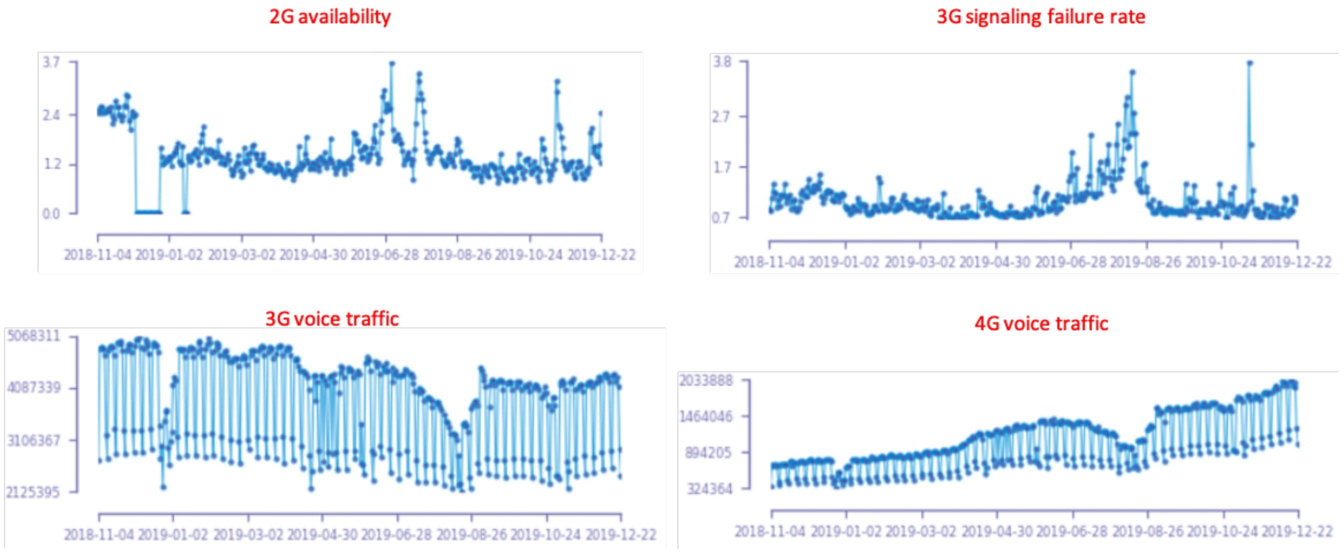

Figure 1. Extract of the training data for four KPIs (in red) over one year. Units of ordinates are pourcentage for top graphs and erlangs for bottom graphs; unit of abscissa is time for all graphs.

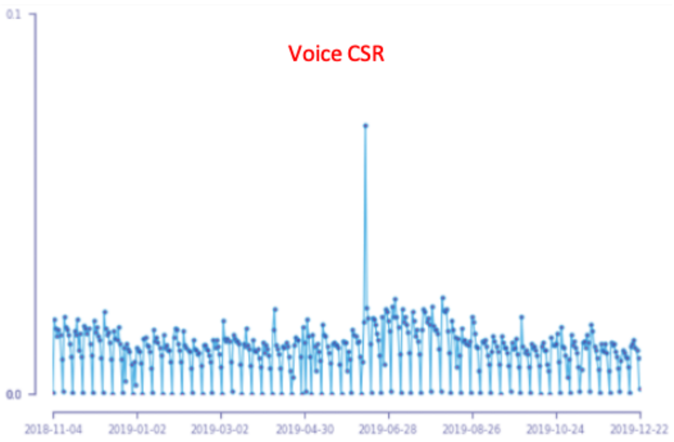

Figure 2. Training data for the voice CSR over one year. Unit of the ordinate is a rate between 0 and 1; unit of the abscissa is time.

B. Least Absolute Shrinkage and Selection Operator

Least Absolute Shrinkage and Selection Operator (LASSO) is a regression method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting model [21]. In other words, the LASSO method handles the complexity of the model with L1 regularization [22], so that the variables not having a contribution to the model are automatically removed from the regression. This means that it adds the “absolute value of magnitude” of coefficients as penalty term to the loss function as shown in Equation 6. LASSO shrinks the less important explanatory variable’s weights to zero thus removing some explanatory variables altogether. This method works well for explanatory variable selection, particularly in case of a huge number of explanatory variables.

$$\begin{aligned}\mathcal{L} &= \min_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{2}\sum_{i=1}^{n}(y^i-\hat{y}^i)^2+\lambda\sum_{j=1}^{p}\mid\beta_j\mid \\ &= \min_{\beta_0,\beta_1,\ldots,\beta_n} \frac{1}{2}\sum_{i=1}^{n}(y^i-\beta_0-\sum_{j=1}^{p}\beta_j x_j^i)^2+\lambda\sum_{j=1}^{p}\mid\beta_j\mid\end{aligned} \tag{6}$$

If λ is set to zero, then LASSO gets back OLS whereas a very large value increases zero coefficients hence it under-fits.

In this work, the fonction `lassocv` of the Python module `statsmodels` has been used to implement LASSO.

V. Data and pre-processing

The goal is to predict the CSR and the influencing factors on a global scale, and not on each specific site, so that performance engineers retrieve aggregated information useful for decision making. The project was hence conducted using data at the level of French *departments* (France has 93 departments that define as many territorial communities) by setting as many regression problems as French departments.

As for the explanatory variables used, the advice of telecom experts led to a mixture of KPIs for both 2G, 3G, and 4G for six classes: traffic (like `downlink data traffic`), availability (like `signaling failure rate`), drop rates, accessibility, performance (like `data_failure rate`), and mobility (like `handover_drop_rate`). In total, 50 KPIs were in the list of explanatory variables, to divide between Data and Voice. Data and Voice are indeed considered to be truly independent from a customer perspective. However, the technical KPIs used by experts to explain voice and data performance have an important common basement. Among the 35 KPIs of the voice list and the 30 KPIs of the data list, 15 KPIs were common to the two lists.

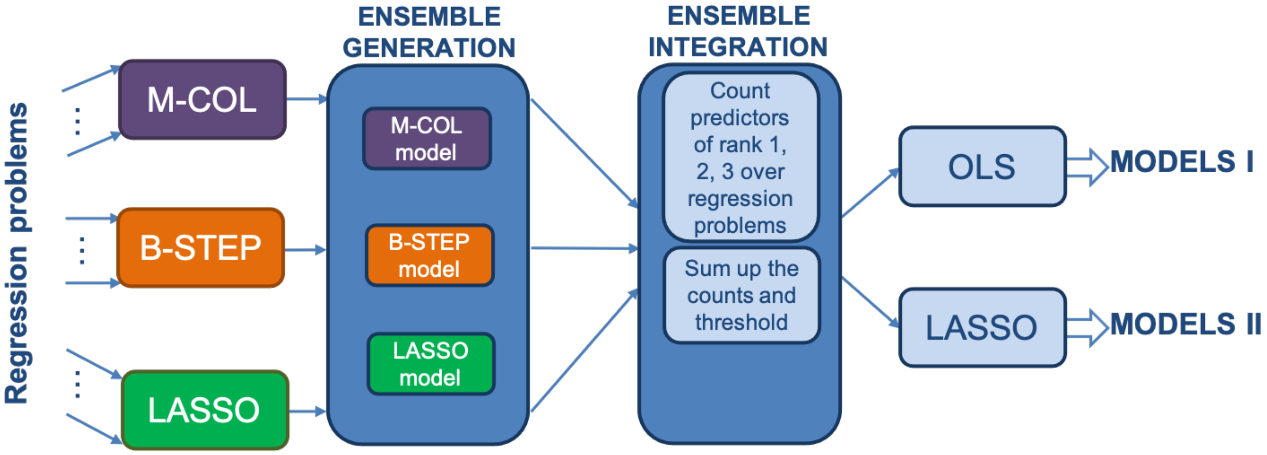

Figure 3. Steps of the fusion regression method

The available data for each department covered a full year. While both daily and weekly values were considered, it was eventually decided to stick with daily ones, to retain a bigger dataset in the training and avoid losing information by averaging over 7 days. An extract of the training data corresponding to four voice KPIs for a specific French department, `2G availability`, `3G signaling failure rate`, `3G voice traffic`, and `4G voice traffic` is shown in Figure 1 on the preceding page. The graph of the corresponding CSR is given in Figure 2 on the previous page.

In a context where the number of explanatory variables is high, it is quite often the case that several variables provide the same information or that some variables remain almost constant, or also that some variables have been poorly sensored. To remedy these common problems, classic data pre-processing solutions were applied in a first step, which consisted in:

- Removing strongly correlated variables, more precisely those with correlation coefficient higher or equal to 0.8;
- Removing variables of low variance through the dataset, more precisely those whose relative standard deviation was lower or equal to 10% of the highest;
- Removing variables with more than 10% missing values. Interpolation was used to fill the gaps for the remaining variables.

In addition, all variables were scaled so that they could be ranked according to the magnitude of their corresponding weights in the regressions.

VI. Ensemble generation

Despite the pre-processing carried out and the elimination of a subset of the KPIs proposed by experts in the field, the number of KPIs remains high, which suggests that still several of them have no direct impact on the CSR. The idea to tackle this problem is to apply an ensemble learning method leveraging the following three base regression approaches, all including a variable selection mechanisms:

- Multicollinearity analysis with OLS (M-COL),
- Backward Stepwise Regression with OLS (B-STEP),
- Structure learning with LASSO (LASSO).

Learning three base regression models with the three methods above constitutes *Step 1* of our ensemble regression method.

Each of the base methods has its own way to tackle the problem of selecting the most relevant explanatory variables, as explained in Sections VI-A, VI-B, and VI-C. To obtain the benefits of the three methods and smooth out the inconsistencies, the three methods are then integrated as explained in Section VII and illustrated in Figure 3. The originality of the proposed ensemble regression integration is that it integrates variable selection instead of directly integrating predictions. This ensemble strategy follows the analysis of [23] whose results suggest the need to examine models using multiple variable selection methods, because when they do not agree, they each may expose different aspects of the complicated theoretical relationships among predictors.

Methods M-COL and B-STEP rely on the classical Ordinary Least Squares method (OLS) presented in Section IV-A whereas LASSO, *Least Absolute Shrinkage and Selection Operator*, uses the method of the same name in its original version of linear regression as presented in Section IV-B.

A. Multicollinearity analysis with OLS

The M-COL method builds on OLS adding an additional preprocessing step that selects a subset of features based on multicollinearity analysis.

In a regression, multicollinearity is a problem that arises when some explanatory variables in the model measure the same phenomenon. Strong multicollinearity is problematic because it can increase the variance of the regression coefficients and make them unstable and difficult to interpret. Strongly correlated predictor coefficients will vary considerably from sample to sample. They may even present the wrong sign.

Multicollinearity does not affect the goodness of the fit or the quality of the forecast. However, the individual coefficients associated with each explanatory variable cannot be interpreted reliably whereas this interpretation is exactly what we are looking for in this work.

Multicollinearity and correlation should not be confused. If collinear variables are de facto strongly correlated with each other, two correlated variables are not necessarily collinear. There is collinearity when two or more variables measure the ”same thing”.

Classically, in case of quantitative explanatory variables, multicollinearity can be assessed by the *variance inflation factor* (VIF) [24]. The VIF for an explanatory variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single explanatory variable. This ratio is calculated for each explanatory variable. The VIF estimates how much the variance of a coefficient is ”increased” due to a linear relationship with other predictors. Thus, a VIF of 1.7 tells us that the variance of this particular coefficient is 70% greater than the variance that should be observed if this factor was absolutely not correlated with the other predictors. The ideal case is obviously when all VIFs are equal to 1, indicating that there is no multicollinearity.

In the case study, multicollinearity analysis was performed considering the 35 and 30 KPIs indicated by the experts in the Voice and Data lists respectively. The VIF threshold was chosen to be 5, beyond which the corresponding KPI was eliminated. Figure 4 shows the results obtained on a specific cell.

B. Backward stepwise regression with OLS

After training a regression model, a *p-value* for each KPI can be obtained: it tests the null hypothesis that the coefficient is equal to zero, in other words, whatever its value, the KPI brings no information whatsoever to the model. A low p-value (typically 0.05 or less) indicates that one can reject the null hypothesis: a predictor that has a low p-value is probably a meaningful addition to the model as it changes the model prediction. Conversely, a larger p-value implies that changes in the predictor do not bring changes in the response.

Backward stepwise selection (or backward elimination) is a variable selection method that begins with a model that contains all variables under consideration (called the Full Model), then removes the least significant variable one after the other based on the p-value until a given stopping condition is satisfied. In our case, the stopping condition states that all remaining variables have a p-value smaller than some pre-specified threshold.

Summarizing, the algorithm is as follows:

- train a model with all KPIs,
- remove the KPI with the highest p-value if it is not lower than a theshold,
- otherwise, stop.

Stepwise regression methods are known to have some drawbacks like instability in the variable selection and biased regression coefficients [25]. However, they may provide efficient means to examine multiple models for further investigation.

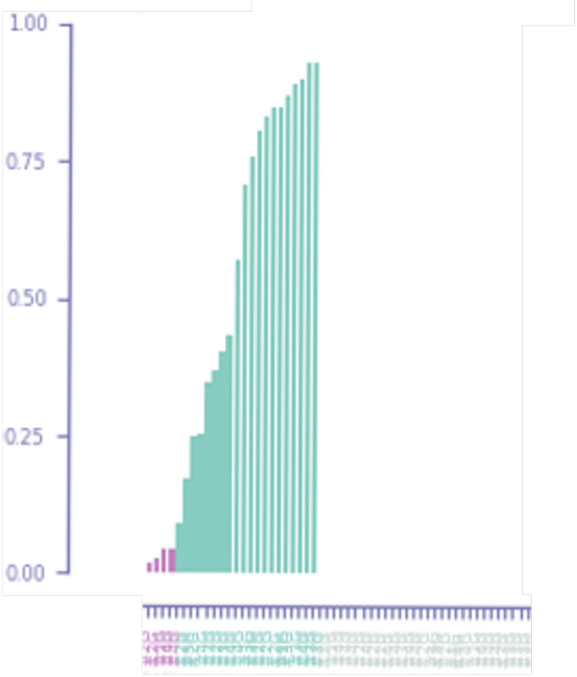

Figure 4. KPI selection and relevancy on a scale from 0 to 1 for a specific cell: grey KPIs are those discarded by pre-processing and multicollinearity analysis, green KPIs are those of minor impact on the CSR, magenta KPIs are those that are preponderant according to the obtained regression model. KPI names have been deliberately blurred.

Note that the problem of biased regression coefficients can be fixed by running a model with the selected variables on a different data set.

C. Structure learning with LASSO

The LASSO method is well known in the literature and has already proved itself in numerous regressions. Here is a quick reminder of the presentation of Section IV-B : in the standard regression like OLS, coefficients are obtained through minimization of the residual squared sum. The LASSO method is similar but adds a penalization term to reduce the number of KPIs kept during the regression. The penalization takes the form of an L1 norm of the coefficients that reduces the available domain of values, allowing some coefficients to be precisely zero, thus letting one remove the matching KPIs.

An advantage of LASSO is that it can be used in high-dimensional problems where the number of observed samples is much smaller than the number of explanatory variables, a case where more classical methods, like OLS, do not work. However, in this very case, if the true vector β is not hollow enough (too many variables of interest), the lasso will not be able to find all these variables of interest. Another limitation is in case of strong correlations, in particular if variables are highly correlated with each other and are important for the prediction, the lasso will favor one of them over the others. Another case, where correlations are a problem, is when the

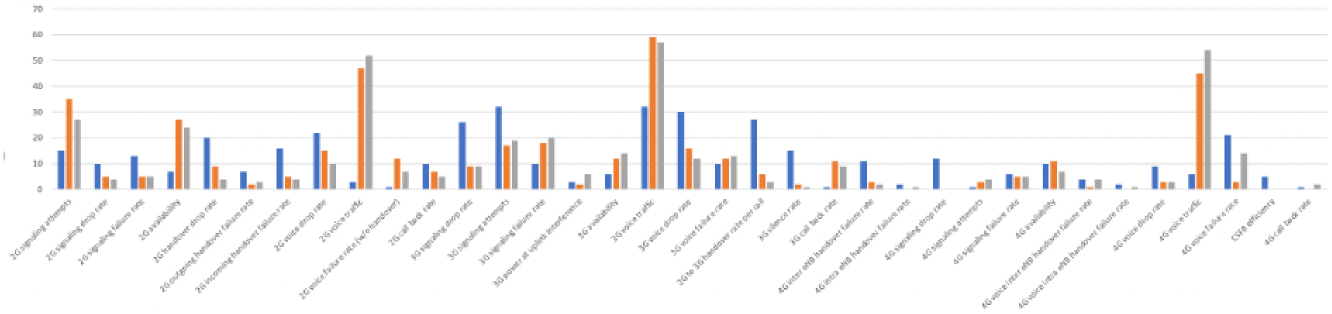

Figure 5. Steps 1-2 of the ensemble integration method for the Voice performance problem: count of the number of times an explanatory KPI is ranked 1, 2, or 3 in the base models from M-COL (blue), B-STEP (orange), and LASSO (grey).

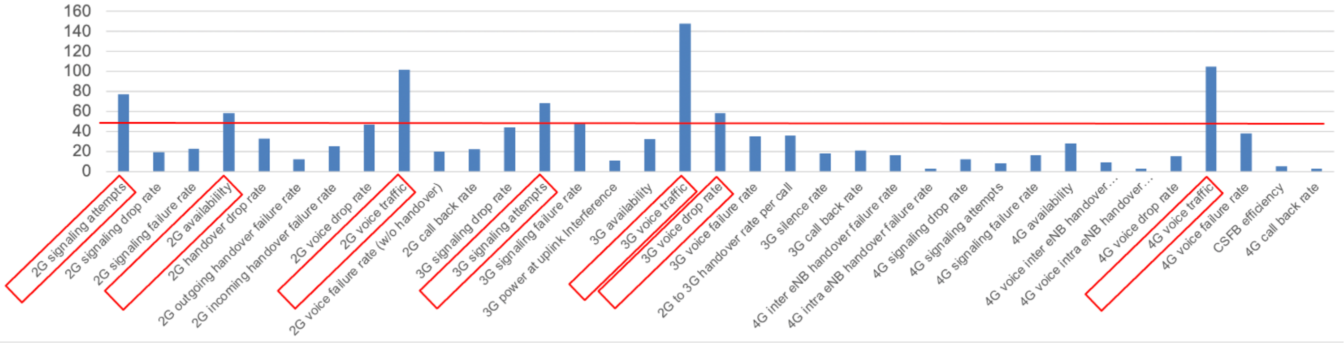

Figure 6. Step 3 of the ensemble integration method for the Voice performance problem: sum of the counts of the number of times an explanatory KPI is ranked 1, 2, or 3 by M-COL, B-STEP, LASSO. KPIs framed in red count above the threshold.

variables of interest are correlated with other variables. In this case, the consistency of the variable selection by LASSO is no longer guaranteed.

VII. Ensemble integration

The principle of ensemble learning is to integrate several learning algorithms to obtain better performance. In this work, ensemble integration is not directly performed on the base model predictions, but on variable selection, for which three base algorithms have been proposed in Section VI. The integrated variable selection is used to learn the final models, hence resulting in indirect regression prediction, as illustrated in Figure 3 on page 5.

Each of the base methods has its own way to tackle the problem of selecting the most relevant explanatory variables, as explained in Sections VI-A, VI-B, and VI-C. Each also comes with a set of advantages and drawbacks.

Ensemble integration aims at obtaining the benefits of the three base algorithms and smooth out their drawbacks, in particular the fact that the base algorithms do not always select the best possible combination of variables.

In the regression model given by (1), explanatory variables $x_j, j = 1, \dots, p$, can be ranked according to the magnitude of their corresponding weight $\beta_1, \dots, \beta_p$. The idea developed in this work uses this ranking and includes four steps for the whole ensemble regression method and three steps for the ensemble integration phase:

- Ensemble generation (as presented in Section VI)
  - *Step 1* – For every regression problem (corresponding to a French department), learn three base regression models with the three selected methods involving explanatory variable selection, namely M-COL, B-STEP, and LASSO;
- Ensemble integration
  - *Step 2* – For M-COL, B-STEP, and LASSO, count the number of times a given explanatory variable (KPI) has rank 1, 2, or 3 over the corresponding base regression models;
  - *Step 3* – Sum up the counts over the three sets of base models and select the explanatory variables whose count exceeds a threshold $\mathcal{T}$;
  - *Step 4* – For every regression problem, learn (on different training data) two integrated regression models with OLS and LASSO considering only the explanatory variables selected at the previous step and deduce the final models and the most impacting variables.

The steps of the ensemble regression method are illustrated in Figure 3 on page 5. The output of the method takes the form of two sets of models called MODELS I and MODELS II,

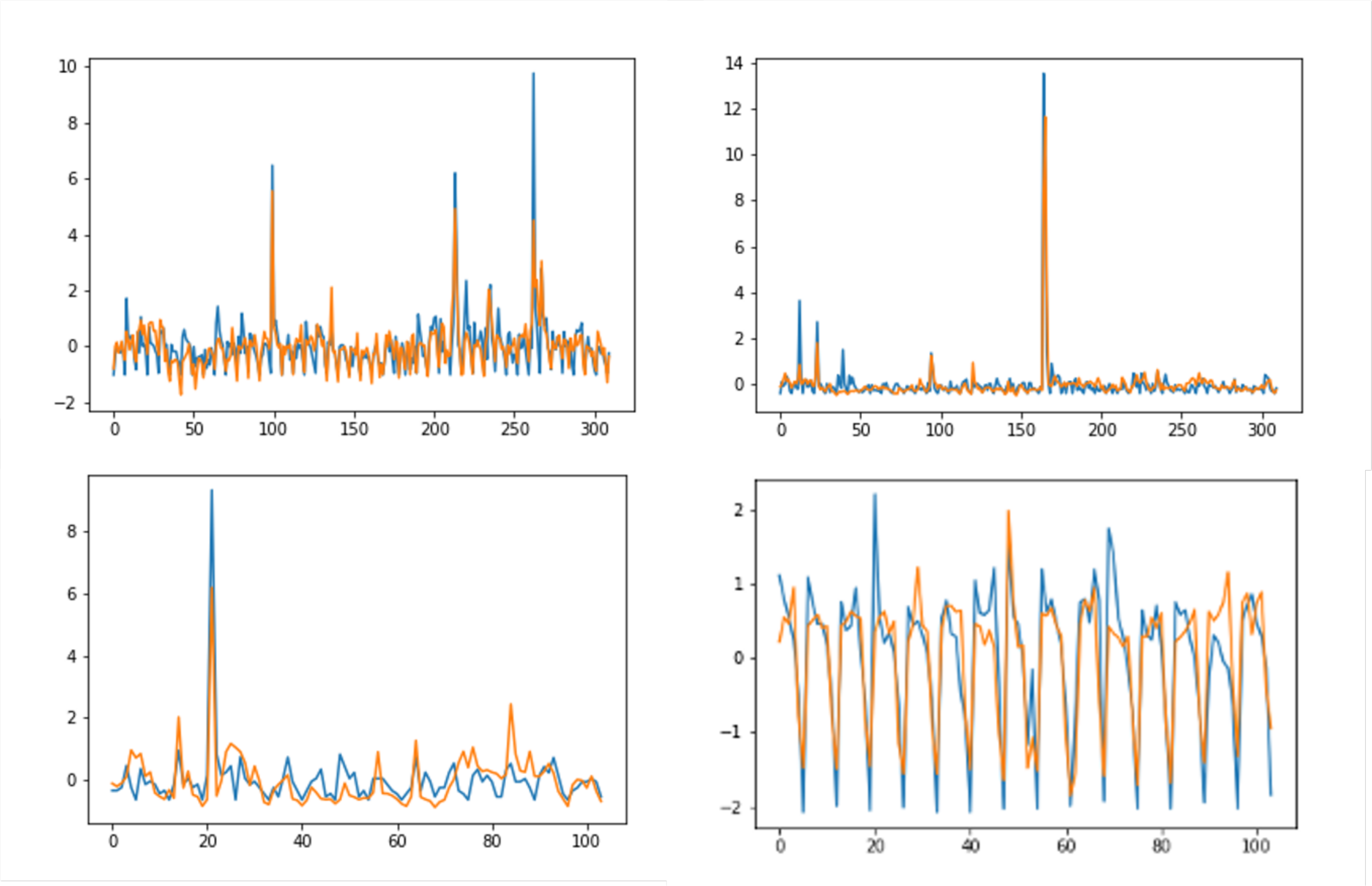

Figure 7. Examples of final models: on training data (top), on test data with larger time scale (bottom).

from which knowledge about most influencing explanatory variables can be extracted as explained in Section VIII.

The ensemble integration method is exemplified with the CSR prediction problems set at the level of French departments.

Steps 1-2 are illustrated in Figure 5 on the preceding page that gives the results for the Voice performance problem. For each explanatory KPI, the blue, orange, and grey bars provide the number of times the KPI is ranked 1, 2 or 3 in the base models obtained by the M-COL, B-STEP, and LASSO method, respectively. Let us note a good convergence of the count referring to B-STEP and LASSO.

Step 3 is illustrated in Figure 6 on the previous page. It aggregates the counts for the base models of each method and sums them up. It hence represents the sum of the counts of the number of times an explanatory KPI is ranked 1, 2, or 3 in the base models obtained by one of the methods M-COL, B-STEP, and LASSO indifferently. A threshold is chosen, here at 45, and the explanatory KPIs that count above this threshold are selected. There are 7 KPIs that count above the threshold, framed in red.

Step 4 considers the 7 ”survivor” KPIs as the most relevant for the prediction of the CSR. This is why step 4 reconsiders every regression problem by restricting explanatory variables to these 7 KPIs. OLS and LASSO methods are run with these explanatory variables alone on another set of training data. Figure 7 shows some examples of the obtained final models on training and test data.

VIII. Making sense of the predictions

Let us recall that the objective of this work is to design a model that makes it possible to link the CSR indicator with a set of objective performance indicators so that performance engineers better understand customer expectations and act first and foremost on the indicators that give the most dissatisfaction. The results of the prediction problems can be analyzed in two ways: at the level of each French department, and aggregated for the whole France.

A. Interpretation at the level of each French department

An interpretation at the level of each French department is done by associating a profile to each department. For this purpose, the results of the final B-STEP models (B-STEP method applied to the 7 survivor KPIs) have been used and the department profiles have been obtained by clustering the coefficients of the obtained models. The clustering was carried

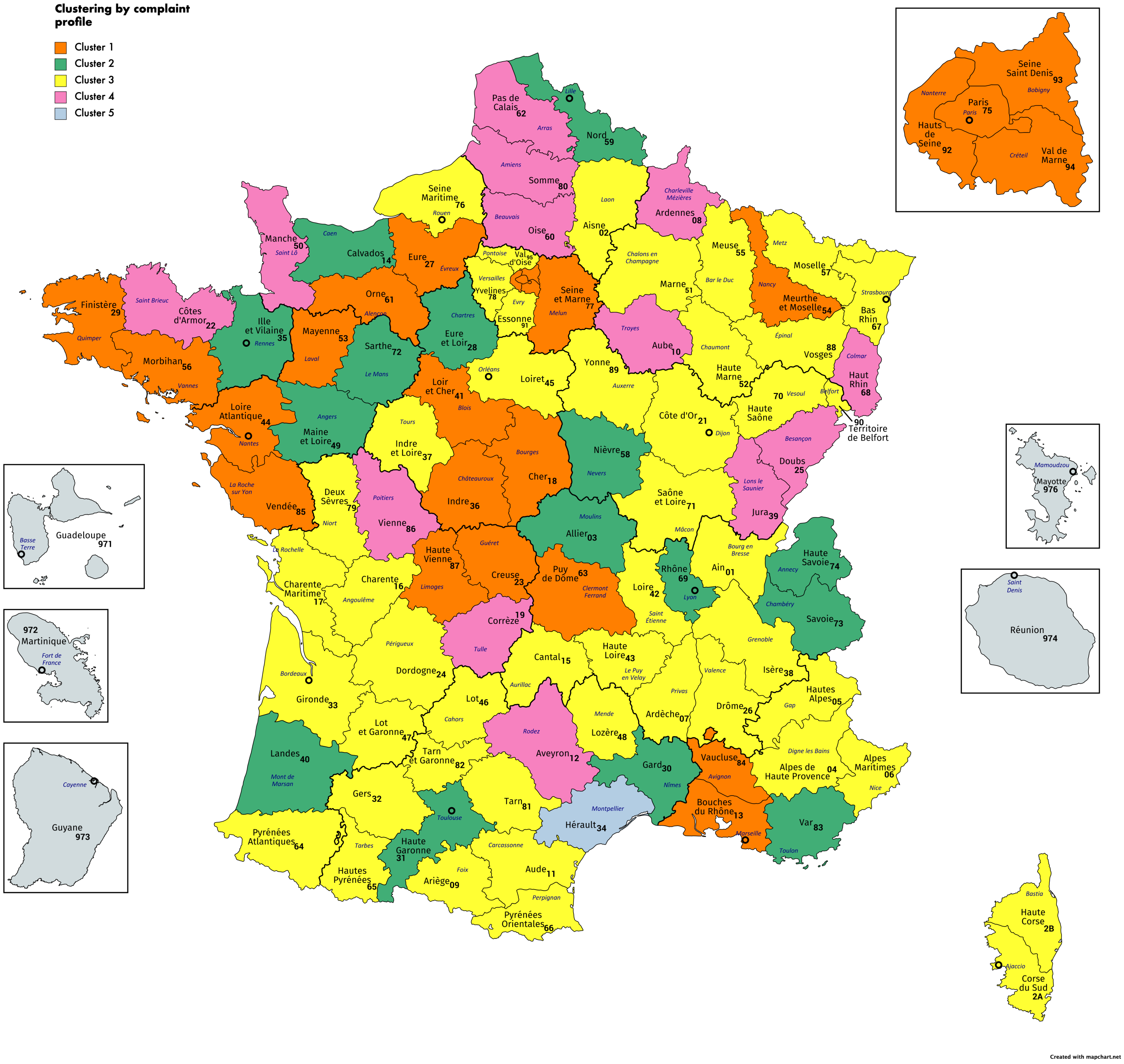

Figure 8. Map of French departments colored by profiles given by the weight of top KPIs influencing the CSR. Departements are identified by their name, number and main city in italics. Overseas departments appear in gray and framed and are not included in the analysis.

out using the classical K-means algorithm that consists in iteratively grouping the individuals (here the models) that are the most similar until stability is reached. Thus, 5 groups emerge whose coefficients associated with each KPI are similar. This leads to the map in Figure 8 where the departments that have similar profile are depicted with the same color. A similar profile indicates that the KPIs that must be mainly incriminated are the same, and so are the reasons explaining customer complaints.

B. Aggregated interpretation

The aggregated interpretation is at the level of the whole France. It requires an additional analysis based on the final models obtained at step 4 of the ensemble integration method. To complete this analysis, KPIs ranked 1, 2, and 3 over OLS and LASSO final models and all the French departments are determined. These KPIs are shown in Figure 9 on the following page, where the three top KPIs (among the 7 survivors) appear in red, namely: `3G voice traffic`, `2G`

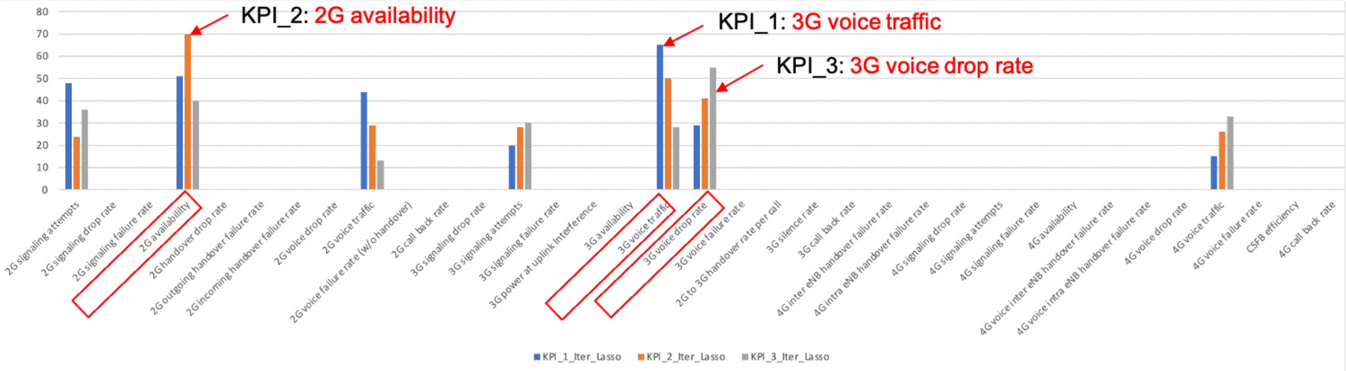

Figure 9. KPIs ranked 1, 2, and 3 over OLS and LASSO final models and over all the French departments.

`availability`, `3G voice drop rate`. These are the most significant KPIs to explain customer dissatisfaction and they indicate that complaints are highly related to network behavior, which is intuitively understandable.

Among the various metrics used to measure network behavior:

- `3G voice traffic` reports about the amount of traffic,
- `2G availability` indicates loss of network coverage,
- `3G voice drop rate` indicates the rate of call drops.

The KPI `3G voice traffic` comes to the first rank. The amount of traffic represented by `3G voice traffic` can be related to network unavailability and network engineering issues. It is easy to understand why these problems may be the main cause of the dissatisfaction of customers.

The KPI `2G availability` comes to the second rank. Loss of network coverage represented by `2G availability` can be associated to network maintenance processes. The fact that this strongly impacts customer dissatisfaction makes sense.

The KPI `3G voice drop rate` comes to the third rank, which is not surprising either.

Let us notice that other metrics like accessibility failure rate or mobility issues appear to be less significant than call drops or traffic issues.

To improve client experience, the network operators should therefore prioritize to base their action plans on:

- reducing unavailability periods by, for instance, optimizing the maintenance process,
- improving the call drop rate by modifying network parameter settings, optimizing site engineering, or building new sites.

IX. Conclusion and Perspectives

This paper proposes an ensemble learning method to obtain a regression model with explanatory power. In many applications, the number of variables that could be thought to be explanatory for a given dependent variable is huge. However, many of them are correlated or collinear and others do not really impact the predicted variable. The method presented in this paper leverages the benefits of three methods to select relevant explanatory variables and deduce a robust regression model. The originality of the ensemble regression integration phase is to focus the integration on variable selection instead of directly on the prediction of the base models.

The method has been tested on telecom data to obtain a model that indicates the impact of a set of objective performance indicators on the customer complaint rate so that performance engineers better understand customer expectations and act first and foremost on the indicators that give the most dissatisfaction. The final results can be used to cluster French departments according to their profile as a function of the top influencing KPIs. Similar profiles indicate that the reasons to be incriminated to explain customer complaints are close, and so are the actions that should be taken. The final results can also be used on a global scale to exhibit the top KPIs at country level and the high level management strategy to be applied.

Future work will consider mapping the top KPIs returned by the model to actual actions to be performed on the network so that customer satisfaction is increased, i.e., CSR is decreased. This mapping could benefit from ideas coming from the combination of the theories of prospect theory and satisfaction games found in the literature, such as [26].

Acknowledgment

We would like to thank the telecom expert Didier Rana for his remarks and relevant suggestions throughout this work.

References

[1] A. Schaffner, L. Travé-Massuyès, S. Pachy, and B. Le Marec, "Explaining radio access network user dissatisfaction with multiple regression models," in *15th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ 2022), Barcelona, Spain*. IARIA, April 2022, pp. 11–17.

[2] J. Jacoby and J. J. Jaccard, "The sources, meaning, and validity of consumer complaint behavior: A psychological analysis." *Journal of Retailing*, vol. 57, no. 3, pp. 4–24, 1981.

[3] J. Singh and R. E. Wilkes, "When consumers complain: A path analysis of the key antecedents of consumer complaint response estimates," *Journal of the Academy of Marketing Science*, vol. 24, no. 4, pp. 350–365, 1996.

[4] J. Goodman and S. Newman, "Understand customer behavior and complaints," *Quality Progress*, vol. 36, no. 1, pp. 51–55, 2003.

[5] W.-H. Au, K. C. Chan, and X. Yao, “A novel evolutionary data mining algorithm with applications to churn prediction,” *IEEE Transactions on Evolutionary Computation*, vol. 7, no. 6, pp. 532–545, 2003.

[6] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–24, 2019.

[7] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, “Customer churn prediction in telecommunication industry using data certainty,” *Journal of Business Research*, vol. 94, pp. 290–301, 2019.

[8] B. Huang, M. T. Kechadi, and B. Buckley, “Customer churn prediction in telecommunications,” *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, 2012.

[9] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, “A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector,” *IEEE Access*, vol. 7, pp. 60 134–60 149, 2019.

[10] Q. Yang, G. Ji, and W. Zhou, “The correlation analysis and prediction between mobile phone users complaints and telecom equipment failures under big data environments,” in *2nd International Conference on Advanced Robotics and Mechatronics (ICARM 2017)*. IEEE, 2017, pp. 201–206.

[11] C. Choi, “Predicting customer complaints in mobile telecom industry using machine learning algorithms,” Ph.D. dissertation, Purdue University, USA, 2018.

[12] R. Polikar, “Ensemble based systems in decision making,” *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[13] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[14] D. Opitz and R. Maclin, “Popular ensemble methods: An empirical study,” *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[15] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, “Ensemble approaches for regression: A survey,” *ACM Computing Surveys (CSUR)*, vol. 45, no. 1, pp. 1–40, 2012.

[16] N. Rooney, D. Patterson, S. Anand, and A. Tsymbal, “Dynamic integration of regression models,” in *International Workshop on Multiple Classifier systems (MCS 2004), Cagliari, Italy*. Springer, June 2004, pp. 164–173.

[17] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[18] G. S. Watson, “Linear least squares regression,” *The Annals of Mathematical Statistics*, pp. 1679–1699, 1967.

[19] I. J. Myung, “Tutorial on maximum likelihood estimation,” *Journal of Mathematical Psychology*, vol. 47, no. 1, pp. 90–100, 2003.

[20] B. Craven and S. M. Islam, “Ordinary least-squares regression,” *The SAGE Dictionary of Quantitative Management Research*, pp. 224–228, 2011.

[21] R. Tibshirani, “Regression shrinkage and selection via the lasso: a retrospective,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 3, pp. 273–282, 2011.

[22] ——, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[23] P. Ruengvirayudh and G. P. Brooks, “Comparing stepwise regression models to the best-subsets models, or, the art of stepwise,” *General Linear Model Journal*, vol. 42, no. 1, pp. 1–14, 2016.

[24] T. A. Craney and J. G. Surles, “Model-dependent variance inflation factor cutoff values,” *Quality Engineering*, vol. 14, no. 3, pp. 391–403, 2002.

[25] G. Smith, “Step away from stepwise,” *Journal of Big Data*, vol. 5, no. 1, pp. 1–12, 2018.

[26] S. Papavassiliou, E. E. Tsiropoulou, P. Promponas, and P. Vamvakas, “A paradigm shift toward satisfaction, realism and efficiency in wireless networks resource sharing,” *IEEE Network*, vol. 35, no. 1, pp. 348–355, 2020.