Colot *et al. J Big Data (2021) 8:63*
https://doi.org/10.1186/s40537-021-00451-9

Journal of Big Data

RESEARCH Open Access

Leveraging fine-grained mobile data for churn detection through Essence Random Forest

Christian Colot[1*], Philippe Baecke[2] and Isabelle Linden[1]

*Correspondence:
christian.colot@unamur.be
[1] Department of Business Administration, University of Namur, Namur, Belgium
Full list of author information is available at the end of the article

Abstract

The rise of unstructured data leads to unprecedented opportunities for marketing applications along with new methodological challenges to leverage such data. In particular, redundancy among the features extracted from this data deserves special attention as it might prevent current methods to benefit from it. In this study, we propose to investigate the value of multiple fine-grained data sources i.e. websurfing, use of applications and geospatial mobility for churn detection within telephone companies. This value is analysed both in substitution and in complement to the value of the well-known communication network. What is more, we also suggest an adaptation of the Random Forest algorithm called Essence Random Forest designed to better address redundancy among extracted features. Analysing fine-grained data of a telephone company, we first find that geo-spatial mobility data might be a good long term alternative to the classical communication network that might become obsolete due to the competition with digital communications. Then, we show that, on the short term, these alternative fine-grained data might complement the communication network for an improved churn detection. In addition, compared to Random Forest and Extremely Randomized Trees, Essence Random Forest better leverages the value of unstructured data by offering an enhanced churn detection regardless of the addressed perspective i.e. substitution or complement. Finally, Essence Random Forest converges faster to stable results which is a salient property in a resource constrained environment.

Keywords: Telecom data, Random Forest, Customer churn, Customer analytics, Unstructured data, Probability models

Introduction

The new era of big data is only at the early stages as digital data is expected to rise from 33 zettabytes in 2018 to 175 zettabytes in 2025 [1]. A large part of this growth will be due to the rise of unstructured data of various forms including data from autonomous vehicles or 5G communications. This evolution is very promising from a marketing perspective as Wedel et al. [2] note that approaches to leverage unstructured data constitute an important issue for research. In particular, they suggest to investigate the combination of geospatial and mobile data with tabular data.

This idea is particularly salient for the specific case of customer relationship management as some studies have already demonstrated the value of fine-grained data for customer acquisition [3] and customer cross-selling [4–6]. However, the paper of Ascarza et al. [7] in 2018 highlights that no study has proven the value of fine-grained data for churn detection beyond communication data records. This research gap is particularly salient for managers of telephone companies as these communication data records are likely to become obsolete due to the competition of digital communications operated by actors such as Messenger or Whats'app [8–10]. Consequently, the ability for telephone companies to detect churn might be deteriorated on the long run. More recently, in 2021, our study [11] showed that referral data might be a good alternative to communication data records to detect churn. This long run opportunity has however two drawbacks. On the one hand, our study highlights that, on the short term, it does not add value on top of communication data records. On the other hand, this strategy also requires starting a potentially expensive referral program, if it is not already the case. In the present study, we investigate the value of multiple alternative internal fine-grained data sources i.e. websurfing, use of applications and geospatial mobility for the following research questions: (RQ1) can these data substitute a communication network on the long run and (RQ2) do these fine-grained data complement a communication network to improve churn detection of a telephone company on the short term? In this investigation, we design multiple feature extraction variations from these data. Unlike more classical data used to make churn prediction such as socio-demographical information or transactional data, there is no standard way to make features which would be guided by previous literature or by domain experts.

This approach of creating multiple features from the same source of data might however increase redundancy in the set of features that are used for modelling churn, which is not desirable as it can lead to suboptimal training of classifiers. Currently, Random Forest can be considered as a state of the art of homogeneous ensemble classifiers [12]. However, this classifier is especially sensitive to redundancy as redundant features are more likely to be randomly chosen to build the trees [13]. While extensive research has been carried out on Random Forest, no proposition tries to directly alleviate this redundancy issue. Consequently, we also propose a variation of this ensemble classifier called Essence Random Forest, which is designed to handle redundancy more efficiently. This proposition leads to the following research questions: (RQ3) does Essence Random Forest offer a better classification performance compared to two benchmark methods of the same family of techniques (namely Random Forest and Extremely Randomized Trees) when dealing with redundant data and (RQ4) does also Essence Random Forest converge faster to a good result as compared to these algorithms ?

The structure of the paper is organised as follows. The second section reports previous studies on fined grained data in churn detection and on Random Forest. The third section discusses our proposition of an adapted Random Forest i.e. Essence Random Forest. The fourth section presents the data of a European telephone company used for this study and the model performance metrics used for this study. The fifth section reports the results of the study in term of value of unstructured mobile data and the contribution of the Essence Random Forest. The sixth section discusses the implications of these results. The last section presents the conclusion of the study and future works.

Related works

The contributions of this paper are twofold: (1) investigating the value of multiple fine-grained data sources for churn detection within telephone companies and (2) proposing a variation of Random Forest to handle the redundancy among the features extracted from these sources. The following subsections discuss existing work related to each contribution.

Fine-grained data for churn detection within telephone companies

According to Ngai et al. [14], churn detection is the customer relationship management (CRM) issue which has received the most attention in the literature. Despite this attention, as noted by Ascarza et al. [7] in 2018, no paper has addressed the value of fine-grained data for churn detection beyond communication data records. Fine-grained data is characterized by high dimensionality and high sparsity meaning, in a human decision context, a small selection of actions among many possibilities. For example, a mobile user can only call a small fraction of other mobile users in the world. As such, communication data records can be considered as fine-grained data. Communication data records are however typically translated into a communication network which can be exploited with the use of social network analysis [15, 16].

More recently, our work [11] addressed the value of two fine-grained types of data for churn detection within telephone companies: referral data and spatial data. Referral data is translated into a referral network where linked referrers and referrals are connected together. Spatial data is converted into a spatial network where edges represent the distance between houses of mobile users. The results show that a referral network does not add value on top of a communication network but it might substitute this network on the long run if this network was to become obsolete due to the competition with digital communications [8–10]. This long term strategy requires however to develop a referral program if it is not already the case.

Compared to this last work, the current article investigates alternative fine-grained data which are already collected by telephone companies in the context of their core activities i.e. web visits, usage of mobile services and geographical mobility. These fine-grained data consequently do not require additional cost to generate and manage it. Further, this investigation might lead to identify fine-grained data which might complement a communication network to detect churn within telephone companies on the short term as it is not the case so far[11]. Finally, from a methodological perspective, these fine-grained data deserve special attention as there is no standard way guided by the literature or domain experts to derive features from this data.

Redundancy issue in Random Forest

To develop churn predictions on these fine-grained data, one has to choose a relevant modelling technique. Random Forest is the current state of the art for classification [12]. This technique has been introduced by Breiman [17] and it is an ensemble method based on the classical decision tree classifier. It aggregates the results of multiple decision trees obtained by randomisation of observations (typically bootstrap) and features. Randomisation of features is obtained by selecting, for each node of the trees, K variables randomly as candidates for split. Both randomisation principles serve as foundations to build decision trees with a minimum correlation between them, which leads to an aggregated result with reduced bias and variance.

This technique is however sensitive to redundancy due to the randomisation of features [13]. For example, cloning several times the same feature increases the probability of this feature of being selected in the candidates set for splitting. This issue is crucial in the current context where multiple feature extractions variations from multiple fine-grained data sources are analysed simultaneously. Some studies propose to perform feature selection before applying the Random Forest classifier to keep only relevant features [18–25]. This approach might however lead to a sub optimal model as less information is provided to the modelling process. Other studies suggest to modify the Random Forest algorithm itself to make it less sensitive to redundancy of features. In particular, several studies researched alternatives for the uniform random sampling of features. Most of them, however, focus on the development of a weighted sampling method based on the bivariate correlation between the target and the feature [26–33]. Overall, the main idea behind these studies suggests to give more weight to features more univariately linked to the target but it does not handle redundancy as this might affect any feature regardless of its univariate link with the target. The study of [34] is an exception to this univariate approach to give weights. It splits the features set into informative features and non informative features based on rough set theory. Then, half of features from each sub-set are selected at random. This approach might lead to exacerbate the issue of redundancy if many uninformative features are redundant. To the best of our knowledge, the study of Kyrillidis et al. [35] is the only study on Random Forest which introduces non uniform weights that are not linked to the association with the target. This study investigates two methods, namely norm-based sampling and statistical leverage scores sampling. While the first method expresses the variance of the feature, the second method uses singular value decomposition to define the weights. These methods give more weight to dominant features but they do not directly handle redundancy as dominant features might be correlated as well.

Outside the investigations on weighted samplings of features, some studies propose to increase the performance of Random Forest by working on feature space, decision criteria to split the node and the aggregation scheme of the trees. Within the works on features space, Rotation Forest is an alternative of Random Forest which fosters a higher accuracy and diversity of base classifiers composing the forest [36]. It creates, for each base classifier, a new set of features derived from Principal Component Analyses performed on a random partition of the feature space into K clusters. This technique may reduce the influence of redundancy in the feature space if features of the same random cluster are correlated. This is however not the case if correlated features fall into different clusters. The work of [37] is a more recent approach in this line of research where the PCA is done on the randomly selected features. This method also aims to foster more diversity of the trees but does not address the redundancy issue that might arise among selected features.

With respect to works on decision criteria, recent research proposes to perform oblique decisions frontiers on the hyperplane available at the node level instead of decisions orthogonal to this hyperplane [38–41]. The main idea consists of using multiple features simultaneously to split the node. This is typically done by fitting models on selected features which might serve as decision frontiers. Some other work in this line of research consists of assigning imprecise probabilities to the labels when measuring the split performance [42, 43] or using random thresholds for the selected features to improve the diversity of trees in the forest [44]. This latter technique called Extremely Randomized Trees is a popular variation of Random Forest (3727 citations) and will be used in our emperical study to provide benchmark results. Lastly, some works suggest alternatives to conventional metrics such as Gini impurity to find the best split criteria based on works in game theory [45, 46]. Overall, these three types of investigation on decision criteria i.e. oblique decision frontiers, increase of randomness or alternative metrics do not specifically handle the redundancy issue as they rely on the random selection of potentially redundant features.

A last line of works consists in investigating the aggregation scheme of the trees. The work of [47] proposes to discard some generated trees of the forest by using a drop out probability. This probability is inversely proportional to the performance of the tree: trees with a weak performance are more likely to be drop out. This procedure aims to select best trees but it does not directly tackle redundancy.

Consequently, to the best of our knowledge, our proposition of Essence Random Forest is the first adaptation of Random Forest which is designed to minimize the influence of redundancy among features.

Essence Random Forest

Our proposition of ensemble classifier designed to be robust to redundancy, called Essence Random Forest, is based on the seminal Random Forest algorithm. The pseudocode to build this benchmark model is expressed in algorithm 1 (partially retrieved from the note of Bernstein [48]). The main function called *RandomForest* (line 1 to 9) performs the randomization of the observations to build each tree and the collection of the resulting tree into the forest. The *RandomizedTreeLearn* function (line 10 to 18) gives more details on the development of each tree including in particular the randomization of features in line 13. As discussed in the related works section, the sensitivity of Random Forest to redundancy is intrinsically linked to this step. By contrast, Essence Random Forest is designed to directly tackle redundancy among features. This ensemble classifier differs from the initial Random Forest by weighting the sampling of features based on the following 2 steps.

Algorithm 1 Build Random Forest

```
Require: A training set S := (x1,y1),...,(xn,yn), features F, and number of trees in forest B
function RandomForest(S, F)
    $H \leftarrow \emptyset$
    for $i \in 1, ..., B$ do
        $S(i) \leftarrow$ A random sample from S
        $hi \leftarrow RandomizedTreeLearn(S(i), F)$
        $H \leftarrow H \cup hi$
    end for
    return H
end function
function RandomizedTreeLearn(S, F)
    while termination conditions not met for every node do
        for every remaining node do
            $f \leftarrow$ simple random subset of F
            Split on best feature in f
        end for
    end while
    return The learned tree
end function
```

First, features are grouped into clusters using a variable clustering technique. In particular, the variable clustering technique used is Varclus [49]. This technique is iterative and contains two phases. The first phase computes the principal component of each cluster based on a principal component analysis. Each feature is then assigned to the component with which it gets the highest squared correlation. The second phase tests if reassigning a feature to another cluster might increase explained variance. Compared to this algorithm, most alternative clustering techniques require to define the number of clusters as input parameter, which is undesirable in our proposition. By contrast, Varclus can also be parametrised by defining the maximum value of the second eigenvalue. The standard threshold value of 0.7 was chosen for this criteria.

Second, weighted probabilities are given to each feature for the random selection of features using the following formula:

$$W_i = \frac{1}{\sum_{j=1}^{NC} NC * C_j * D_{i,j}} \tag{1}$$

where W_i is the weight for feature X_i, NC is the number of clusters, C_j is the number of features in cluster j and $D_{i,j}$ is a dummy variable which takes value 1 if X_i belongs to cluster j, 0 otherwise.

Algorithm 2 Build Essence Random Forest

```
Require: A training set S := (x1,y1),...,(xn,yn), features F, and number of trees in forest B
 1: function EssenceRandomForest(S, F)
 2:     Group features into clusters of correlated variables
 3:     Compute W, the vector of weights based on formula 1
 4:     $H \leftarrow \emptyset$
 5:     for $i \in 1, ..., B$ do
 6:         $S(i) \leftarrow$ A random sample from S
 7:         $hi \leftarrow EssenceRandomizedTreeLearn(S(i), F)$
 8:         $H \leftarrow H \cup hi$
 9:     end for
10:     return H
11: end function
12: function EssenceRandomizedTreeLearn(S, F)
13:     while termination conditions not met for every node do
14:         for every remaining node do
15:             $f \leftarrow$ random subset of F using W as distribution probability
16:             Split on best feature in f
17:         end for
18:     end while
19:     return The learned tree
20: end function
```

The pseudo-code of Essence Random Forest is given in algorithm 2[1]. The first difference compared to the original algorithm (i.e. building clusters and deriving probabilities) is expressed in line 2 and 3, while the second modification (i.e. using probabilities for random selection) is defined in line 15 of the new algorithm. Using this method, the effect of cloned or very similar features is neutralised. To illustrate this property, consider the following pedagogical example. For a data modelling context, there are only two features available (X_1 and X_2) and K = 1 meaning only one feature selected at random for each node to split. In this case, both Random Forest and Essence Random Forest give 50% of chance for each feature of being selected. Now, X_2 is duplicated 98 times leading to a set of 100 features. In this new case, two clusters will be created: 1 with 1 variable and 1 with 99 variables. Next applying, formula 1, X_1 will have a 50% to be selected and each X_2 duplicate will have a 0.5051% chance to be selected which sums up to also 50% overall.

Let us now discuss the implications of these modifications in term of performance of the algorithm. As discussed by [17], the generalization error of Random Forest has the following upper bound:

$$PE \leq \frac{\bar{\rho}\left(1 - s^2\right)}{s^2} \tag{2}$$

[1] The source code in Python is available on Github at the following link: https://github.com/christiancolot/EssenceRandomForest.

Table 1 Base and communication metrics

Category	Feature
Base	Age
	Gender
	Time from enrolment
	Average evolution of data volume last month compared to previous month
	Average evolution of number of text messages last month compared to previous month
	Average evolution of total number of minutes of outgoing calls last month compared to previous month
Communication network	Number of connections (i.e. degree)
	Average churn behavior among peers weighted by communication
	Total communication time spent with defecting peers

where $\bar{\rho}$ is the average correlation between trees and s^2 the mean performance of each individual tree. Consequently, the performance of Random Forest is driven by the ability of individual trees and by the diversity of the trees. Essence Random Forest refines both of these ingredients because Essence Random Forest performs the random selection at the level of the information and not at the level of the feature. In case of redundancy, Essence Random Forest fosters the possibility to still build trees with a diversity of features in terms of information. This diversity captured in the random process of selecting features allows capturing more information within a single tree and increasing the heterogeneity of information between the trees.

Methodology

Data description and featurization

As mentioned earlier, this study has two main contributions: testing the value of multiple internal fine-grained data sources for churn detection within telephone companies and, in this context, assessing the value of an Essence Random Forest to handle redundancy. For this, data from an European telephone company was collected in 2017. The data contains information on customer socio-demographics, transactions, call data records, web browsing, usage of services and geographical mobility. The sample contains 631 619 retail postpaid customers with regular phone usage. The predictive model will predict whether these customers are likely to leave the telephone company within the next two months. Five months of data are used to build the model. The first month is dedicated to build the communication network among customers and the bipartite graphs from fine-grained data that will be described next. The second and third months are allocated to build features including the churn behaviour of customers peers. The two last months are used to assess the churn behaviours of the focal customers.

From the data, 87 features are extracted. These features are categorised into five categories: base metrics, communication metrics, url metrics, service metrics and geo-spatial metrics, depending on the data source. Base metrics contains socio-demographical information and activity variables coming from the CRM system of the telephone company. These metrics are traditionally used in churn detection models (see Table 1).

Communication metrics are retrieved from call data records (CDRs). CDRs are a fine-grained data source in which each communication activity (e.g. call, text message) between the customer and somebody else is stored. This data source can be leveraged by drawing a communication network from these communication activities. In this network, customers are represented as nodes and communication intensities as weights of the edges. Following Ma et al. [50], two customers are connected in this network if there is a bidirectional communication and if there are at least 5 communications among them taking place during the first month of the data. These conditions are designed to only connect in the network customers who have regular reciprocal contacts so that they might be subject to peer influence. This represents the fact that two connected individuals might exhibit more regularly the same behaviour than two disconnected individuals. There are two main sources of peer influence: homophily i.e. the fact of sharing some similarity traits like the same level of education or income [51] and social influence [52]. Three metrics were retrieved from this network (see Table 1) to express this influence.

Url metrics are derived from another fine-grained data source that collects the web browsing of the user on a daily granularity level. In particular, this data source does not only contain the visit of the website but also some other raw variables: total visits, total visits duration, total pages viewed, aggregation level 1 of website and aggregation level 2 of website. These aggregations levels are based on the internal classification of websites of the telephone company. A basic approach in which traditional metrics are calculated for every website would lead to an extremely high dimensional sparse data which would not be appropriate for modelling. To featurize this data in a more convenient way, we rely on the works of Stankova et al. [53] and De Montjoye et al. [54]. Stankova et al. [53] investigated how bipartite graphs can be used for classification purposes. Visits of customers to websites can be formulated in bipartite graphs where customers and websites represent two types of nodes. Inspired by the SW-transformation approach that they propose to get a scalable featurization process, we propose the following general formula:

$$UrlAvgChurn_{j,k,l} = \frac{\sum_{s \varepsilon N(i,j)} \frac{IM_{i,s,k}}{M_{s,k} * D_l} * \frac{MC_{s,k}}{M_{s,k}}}{\sum_{s \varepsilon N(i,j)} \frac{IM_{i,s,k}}{M_{s,k} * D_l}} \quad (3)$$

where i represents the focal customer, j the level of granularity for the website dimension, k the raw metric(i.e. visit, total visits, total visits duration, total pages viewed), l the focus level (overall users or individual user), s a member of the website dimension (i.e. website or corresponding aggregation level), $IM_{i,s,k}$ the individual contribution of customer i for member s concerning metric k, $M_{s,k}$ the sum of all individual contributions of users of member s concerning metric k, $MC_{s,k}$ the sum of all individual contributions of churners of member s concerning metric k, N(i,j) the subspace including members of the level j visited by customer i and D_l a dummy variable where $D_1 = 1$ and $D_2 = 1/M_{s,k}$.

Compared to the formula proposed by Stankova et al. [53], our formula computes a probability and not a sum proportional to this probability. This change is important as, in their formula, deriving a probability would require to draw an univariate projection of the bigraph which would lead to withdraw the scalability property. Consequently, our formula does not tend to give higher value for customers visiting more websites. What is more, our formula generalises to the possibility of integrating multiple raw metrics, multiple aggregation levels for the website dimension and multiple perspectives that we explain hereafter.

This formula contains two main components: a weight i.e. $\frac{IM_{i,s,k}}{M_{s,k}*D_l}$ and a proportion $\frac{MC_{s,k}}{M_{s,k}}$. The weight can be expressed on the overall level or individual level depending on the value of l. On the overall level, the weight expresses the individual contribution of the user compared to the overall activity on a given website granularity level. If the overall activity is high, this weight tends to be lower. Following Stankova et al. [53], this expresses the idea that mainstream websites would be less indicative of individual behaviours contrary to websites with less activity which tend to give more information on the user. This effect is moderated by the activity of the customer itselves: the more he or she contributes to this overall activity, the more the website is likely to reflect his or her individual behaviour hence an higher weight. This is not the case for the study of Stankova et al. [53] where every customer gets the same weight from a website (i.e. top node) whatever his or her individual contribution. On the individual level, the overall activity is even neutralised meaning that we do only take into account the personal behaviour of the customer: a website (or an corresponding aggregation level) will be considered important if his or her activity on this website is important in his or her global activity.

The proportion is the activity of churners for the given metric k compared to the activity of all users. The weighted average of this proportion across all websites (or a corresponding aggregation level) gives a value between 0 and 1. It can consequently be considered as a probability. Combining the three granularity levels of url, the four metrics and the two focus levels lead to 24 url metrics based on formula 3. See Table 2.

These metrics are based on the churn behaviour of peers visiting the same website. We also retrieved features of url source which do not directly include this behaviour. In particular, these metrics are based on the work of De Montjoye et al. [54]. Based on a quasi experimental design, they recorded the mobile phone activity of 69 people during 4 months. At the same time, the participants had also to fill in a personality questionnaire using the five factor model [55]. Based on a literature review in personality psychology, they defined features extracted from mobile data to predict personality traits. They found a 42% increase in accuracy compared to random. Four metrics were adjusted from this study to apply to the case of websites: entropy duration of website visits, entropy visits of website, number of unique websites and total visits to number of unique websites ratio. The entropy measures the extent to which each individual website gets the same attention. For each metric, the three granularity levels were considered, leading to 12 additional metrics. The underlying idea is that personality traits of the customer disclosed by url patterns might influence the customer's churn behaviour.

Service metrics are derived from another fine-grained data source. This data source collects the usage of mobile services such as applications on a daily granularity level. These metrics are computed in exactly the same way as url metrics leading to 24 metrics linked to churn activity of customers with the same service usage pattern and 12 metrics linked to personality traits of the user. The only notable differences are the available raw metrics which are in this case: total data volume, session count and total session duration. See Table 3.

Geo-spatial metrics are derived from a data source that tracks the GPS coordinates of the nearest cell tower when communicating by text messages or call. The same

Table 2 Url metrics

Feature	Label
Entropy duration of website visits—level 1	Entropy duration of website visits—level 1
Entropy visits of website—level 1	Entropy visits of website—level 1
Number of unique websites—level 1	Number of unique websites—level 1
Total visits to number of unique websites ratio—level 1	Total visits to number of unique websites ratio—level 1
Entropy duration of website visits—level 2	Entropy duration of website visits—level 2
Entropy visits of website—level 2	Entropy visits of website—level 2
Number of unique websites—level 2	Number of unique websites—level 2
Total visits to number of unique websites ratio—level 2	Total visits to number of unique websites ratio—level 2
Entropy duration of website visits—level 3	Entropy duration of website visits—level 3
Entropy visits of website—level 3	Entropy visits of website—level 3
Number of unique websites—level 3	Number of unique websites—level 3
Total visits to number of unique websites ratio—level 3	Total visits to number of unique websites ratio—level 3
$\text{UrlAvgChurn}_{\text{level1,visits,overall}}$	Churn % of people visiting same url level1 weighted by contribution to overall visits
$\text{UrlAvgChurn}_{\text{level1,totalvisits,overall}}$	Churn % of people visiting same url level1 weighted by contribution to overall total visits
$\text{UrlAvgChurn}_{\text{level1,totalvisitsduration,overall}}$	Churn % of people visiting same url level1 weighted by contribution to overall total visits duration
$\text{UrlAvgChurn}_{\text{level1,totalpagesviewed,overall}}$	Churn % of people visiting same url level1 weighted by contribution to overall total pages viewed
$\text{UrlAvgChurn}_{\text{level1,visits,individual}}$	Churn % of people visiting same url level1 weighted by contribution to individual visits
$\text{UrlAvgChurn}_{\text{level1,totalvisits,individual}}$	Churn% of people visiting same url level1 weighted by contribution to individual total visits
$\text{UrlAvgChurn}_{\text{level1,totalvisitsduration,individual}}$	Churn % of people visiting same url level1 weighted by contribution to individual total visits duration
$\text{UrlAvgChurn}_{\text{level1,totalpagesviewed,individual}}$	Churn % of people visiting same url level1 weighted by contribution to individual total pages viewed
$\text{UrlAvgChurn}_{\text{level2,visits,overall}}$	Churn % of people visiting same url level2 weighted by contribution to overall visits
$\text{UrlAvgChu}_{\text{level2,totalvisits,overall}}$	Churn % of people visiting same url level2 weighted by contribution to overall total visits
$\text{UrlAvgChurn}_{\text{level2,totalvisitsduration,overall}}$	Churn % of people visiting same url level2 weighted by contribution to overall total visits duration
$\text{UrlAvgChurn}_{\text{level2,totalpagesviewed,overall}}$	Churn % of people visiting same url level2 weighted by contribution to overall total pages viewed
$\text{UrlAvgChurn}_{\text{level2,visits,individual}}$	Churn % of people visiting same url level2 weighted by contribution to individual visits
$\text{UrlAvgChurn}_{\text{level2,totalvisits,individual}}$	Churn % of people visiting same url level2 weighted by contribution to individual total visits
$\text{UrlAvgChurn}_{\text{level2,totalvisitsduration,individual}}$	Churn % of people visiting same url level2 weighted by contribution to individual total visits duration
$\text{UrlAvgChurn}_{\text{level2,totalpagesviewed,individual}}$	Churn % of people visiting same url level2 weighted by contribution to individual total pages viewed
$\text{UrlAvgChurn}_{\text{level3,visits,overall}}$	Churn % of people visiting same url level3 weighted by contribution to overall visits
$\text{UrlAvgChurn}_{\text{level3,totalvisits,overall}}$	Churn % of people visiting same url level3 weighted by contribution to overall total visits
$\text{UrlAvgChurn}_{\text{level3,totalvisitsduration,overall}}$	Churn % of people visiting same url level3 weighted by contribution to overall total visits duration
$\text{UrlAvgChurn}_{\text{level3,totalpagesviewed,overall}}$	Churn % of people visiting same url level3 weighted by contribution to overall total pages viewed
$\text{UrlAvgChurn}_{\text{level3,visits,individual}}$	Churn % of people visiting same url level3 weighted by contribution to individual visits

Table 2 (continued)

Feature	Label
$\mathrm{UrlAvgChurn}_{\mathrm{level3,totalvisits,individual}}$	Churn % of people visiting same url level3 weighted by contribution to individual total visits
$\mathrm{UrlAvgChurn}_{\mathrm{level3,totalvisitsduration,individual}}$	Churn % of people visiting same url level3 weighted by contribution to individual total visits duration
$\mathrm{UrlAvgChurn}_{\mathrm{level3,totalpagesviewed.individual}}$	Churn % of people visiting same url level3 weighted by contribution to individual total pages viewed

equation as 3 is applied. In particular, there is here one granularity level, one raw metric (visit of a GPS coordinates) and 2 perspectives (individual or overall), leading to 2 geo-spatial metrics. This category also includes 4 metrics linked to personality traits. See Table 4.

Modelling design and evaluation metrics

From the previous sub-section, we can observe that some features might be correlated in particular among variations of Eq. 3. This is why we investigate not only the value of fine-grained data sources but also how to handle the redundancy that comes with it. In our modelling design, we compare the results obtained with our variation of Random Forest designed to handle redundancy called Essence Random Forest to the results of two methods of the same family of techniques namely Random Forest [17] and Extremely Randomized Trees [44]. This last method, discussed in the related works as a work on the refinement of the decision criteria, is a popular technique (3727 citations). Compared to Essence Random Forest, both techniques aim to promote more diversity between the trees than in the seminal Random Forest method. However, the random choice of the cut-off in Extremely Randomized Trees may degrade the performance of a single tree, while Essence Random Forest fosters a better performance of each tree by considering more diversity of information to choose the best feature to split a node. The modelling design will also compare multiple levels of number of trees in the ensemble model to compare to which extent each technique quickly converges to a stable performance. Data for modelling is randomly split with stratification on the target between two sets.

The first set, including 30% of the data, is used to fine tune hyper parameters specific to this kind of ensemble method: max depth of the tree (from 1 to 10), minimum number of observations required to split a node (2, 10, 20, 30, 40, 50 and 60), minimum number of observations required in a final node (1, 5, 10, 20, 30) and sampling of observations (bootstrap or not). For the number of randomly selected features, the standard value i.e. square root of the number of features is chosen as suggested by Breiman [56]. As discussed above, the number of trees is also not part of this parameter tuning as we want to assess the influence of this parameter on the results. To find a good value for all tuned parameters, a ten-fold randomized parameter search was applied. Compared to a grid search, the randomized search does not cover all possible combinations of the values (700 in our setting) but randomly chooses some of them. A value of 100 combinations is kept for the analysis. This randomized search is used to get good results in an affordable computer time (almost one month of computer time needed with this last

Table 3 Service metrics

Feature	Label
Entropy duration of service visits—level 1	Entropy duration of service visits—level 1
Entropy visits of services—level 1	Entropy visits of services—level 1
Number of unique services—level 1	Number of unique services—level 1
Total data volume to number of unique services ratio—level 1	Total data volume to number of unique servicesratio—level 1
Entropy duration of service visits—level 2	Entropy duration of service visits—level 2
Entropy visits of services—level 2	Entropy visits of services—level 2
Number of unique services—level 2	Number of unique services—level 2
Total data volume to number of unique services ratio—level 2	Total data volume to number of unique servicesratio—level 2
entropy duration of service visits—level 3	Entropy duration of service visits—level 3
Entropy visits of services—level 3	Entropy visits of services—level 3
Number of unique services—level 3	Number of unique services—level 3
Total data volume to number of unique services ratio—level 3	Total data volume to number of unique servicesratio—level 3
$\text{ServiceAvgChurn}_{\text{level1,visits,overall}}$	Churn % of people visiting same service level1 weighted by contribution to overall visits
$\text{ServiceAvgChurn}_{\text{level1,totaldatavolume,overall}}$	Churn % of people visiting same service level1 weighted by contribution to overall total data volume
$\text{ServiceAvgChurn}_{\text{level1,sessioncount,overall}}$	Churn % of people visiting same service level1 weighted by contribution to overall session count
$\text{ServiceAvgChurn}_{\text{level1,totalsessionduration,overall}}$	Churn % of people visiting same service level1 weighted by contribution to overall total session duration
$\text{ServiceAvgChurn}_{\text{level1,visits,individual}}$	Churn % of people visiting same service level1 weighted by contribution to individual visits
$\text{ServiceAvgChurn}_{\text{level1,totaldatavolume,individual}}$	Churn % of people visiting same service level1 weighted by contribution to individual total data volume
$\text{ServiceAvgChurn}_{\text{level1,sessioncount,individual}}$	Churn % of people visiting same service level1 weighted by contribution to individual session count
$\text{ServiceAvgChurn}_{\text{level1,totalsessionduration,individual}}$	Churn % of people visiting same service level1 weighted by contribution to individual total session duration
$\text{ServiceAvgChurn}_{\text{level2,visits,overall}}$	Churn % of people visiting same service level2 weighted by contribution to overall visits
$\text{ServiceAvgChurn}_{\text{level2,totaldatavolume,overall}}$	Churn % of people visiting same service level2 weighted by contribution to overall total data volume
$\text{ServiceAvgChurn}_{\text{level2,sessioncount,overall}}$	Churn % of people visiting same service level2 weighted by contribution to overall session count
$\text{ServiceAvgChurn}_{\text{level2,totalsessionduration,overall}}$	Churn % of people visiting same service level2 weighted by contribution to overall total session duration
$\text{ServiceAvgChurn}_{\text{level2,visits,individual}}$	Churn % of people visiting same service level2 weighted by contribution to individual visits
$\text{ServiceAvgChurn}_{\text{level2,totaldatavolume,individual}}$	Churn % of people visiting same service level2 weighted by contribution to individual total data volume
$\text{ServiceAvgChurn}_{\text{level2,sessioncount,individual}}$	Churn % of people visiting same service level2 weighted by contribution to individual session count
$\text{ServiceAvgChurn}_{\text{level2,totalsessionduration,individual}}$	Churn % of people visiting same service level2 weighted by contribution to individual total session duration
$\text{ServiceAvgChurn}_{\text{level3,visits,overall}}$	Churn % of people visiting same service level3 weighted by contribution to overall visits

Table 3 (continued)

Feature	Label
$\text{ServiceAvgChurn}_{\text{level3,totaldatavolume,overall}}$	Churn % of people visiting same service level3 weighted by contribution to overall total data volume
$\text{ServiceAvgChurn}_{\text{level3,sessioncount,overall}}$	Churn % of people visiting same service level3 weighted by contribution to overall session count
$\text{ServiceAvgChurn}_{\text{level3,totalsessionduration,overall}}$	Churn % of people visiting same service level3 weighted by contribution to overall total session duration
$\text{ServiceAvgChurn}_{\text{level3,visits,individual}}$	Churn % of people visiting same service level3 weighted by contribution to individual visits
$\text{ServiceAvgChurn}_{\text{level3,totaldatavolume,individual}}$	Churn % of people visiting same service level3 weighted by contribution to individual total data volume
$\text{ServiceAvgChurn}_{\text{level3,sessioncount,individual}}$	Churn % of people visiting same service level3 weighted by contribution to individual session count
$\text{ServiceAvgChurn}_{\text{level3,totalsessionduration,individual}}$	Churn % of people visiting same service level3 weighted by contribution to individual total session duration

Table 4 Geo-spatial metrics

Feature
Average daily radius of gyration Average daily distance traveled
Entropy of places visited: large values indicate regular visits to many places
$\text{GeoAvgChurn}_{\text{level1,totalvisits,overall}}$: churn % of people visiting same areas weighted by own contribution to overall visits
$\text{GeoAvgChurn}_{\text{level1,totalvisits,individual}}$: churn % of people visiting same areas weighted by number of own visits
Number of places visited

setting instead of 7 months on a I7-7800X CPU @ 3.5 GHZ including 6 cores, 12 logical cores and 64 GB Ram). The best combination identified is the one with the best average performance on our main performance metric i.e. area under the roc curve (AUC) that we will describe below.

The second set, including 70% of the data, is used to evaluate the performance of the model. A 10-fold cross-validation process is performed with the parameters values identified with the first step. The average AUC on the 10 test set is then computed. The AUC is a metric which does not require choosing a specific threshold. It can be interpreted as the probability that the model identifies the right churner among two randomly chosen customers, one of the two being selected in the churner set. To assess the statistical significance of the AUC difference between two models, the Delong test is used [57]. Finally, to further interpret the contribution of each feature to a specific model, variable importance metrics are computed [17].

Results

Table 5 reports AUC results for Random Forest (RF), Extremely Randomized Trees (ERT) and Essence Random Forest (ERF) for different levels of number of trees and different models. Six models are considered, from a benchmark model including only base

Table 5 Results: AUC performance for Random Forest (RF), Extremely Randomized Trees (ERT) and Essence Random Forest (ERF)

Nb	Base			Base + Com			Base + Geo			Base + Url			Base + Service			Full		
trees	RF	ERT	ERF	RF	ERT	ERF	RF	ERT	ERF	RF	ERT	ERF	RF	ERT	ERF	RF	ERT	ERF
10	61.05%	60.90%	60.93%	61.41%	61.16%	61.51%	61.71%	61.36%	61.83%	60.85%	59.18%	61.40%	61.08%	60.26%	61.66%	61.20%	60.30%	62.38%
20	61.07%	61.04%	61.06%	61.65%	61.45%	61.66%	61,99%	61.63%	62.03%	61.17%	60.64%	61.60%	61.51%	60.84%	61.73%	62.07%	60.63%	62.60%
30	61.12%	61.07%	61.17%	61.61%	61.53%	61.63%	62,04%	61.74%	62.03%	61.28%	60.54%	61.57%	61.55%	61.03%	61.95%	62.11%	61.08%	62.65%
40	61.13%	60.99%	61.20%	61.64%	61.53%	61.65%	62,11%	61.78%	62.01%	61.45%	60.51%	61.54%	61.59%	60.84%	61.76%	62.33%	61.06%	62.68%
50	61.20%	61.02%	61.17%	61.72%	61.52%	61.71%	62,10%	61.86%	62.01%	61.35%	60.64%	61.66%	61.69%	60.95%	61.90%	62.38%	61.12%	62.71%
60	61.19%	61.07%	61.19%	61.65%	61.54%	61.69%	62,12%	61.83%	62.01%	61.48%	60.76%	61.65%	61.63%	61.11%	61.82%	62.39%	61.24%	62.73%
70	61.19%	61.09%	61.21%	61.71%	61.52%	61.70%	62,15%	61.81%	62.07%	61.35%	60.97%	61.69%	61.69%	61.03%	61.90%	62.37%	61.18%	62.81%
80	61.21%	61.12%	61.22%	61.67%	61.57%	61.76%	62,13%	61.79%	62.16%	61.42%	60.84%	61.66%	61.71%	61.12%	61.92%	62.49%	61.72%	62.74%
90	61.22%	61.09%	61.23%	61.69%	61.50%	61.74%	62,17%	61.91%	62.17%	61.49%	60.77%	61.66%	61.74%	61.15%	61.86%	62.37%	61.79%	62.81%
100	61.22%	61.12%	61.23%	61.71%	61.50%	61.75%	62,24%	61.89%	62.08%	61.46%	60.85%	61.69%	61.75%	60.98%	61.92%	62.55%	61.86%	62.87%
200	61.30%	61.12%	61.29%	61.74%	61.54%	61.76%	62,12%	61.85%	62.19%	61.52%	60.96%	61.72%	61.72%	61.27%	62.01%	62.60%	61.97%	62.91%
500	61.29%	61.17%	61.26%	61.75%	61.57%	61.77%	62.29%	61.97%	62.24%	61.55%	61.02%	61.73%	61.77%	61.26%	62.00%	62.56%	62.08%	62.94%

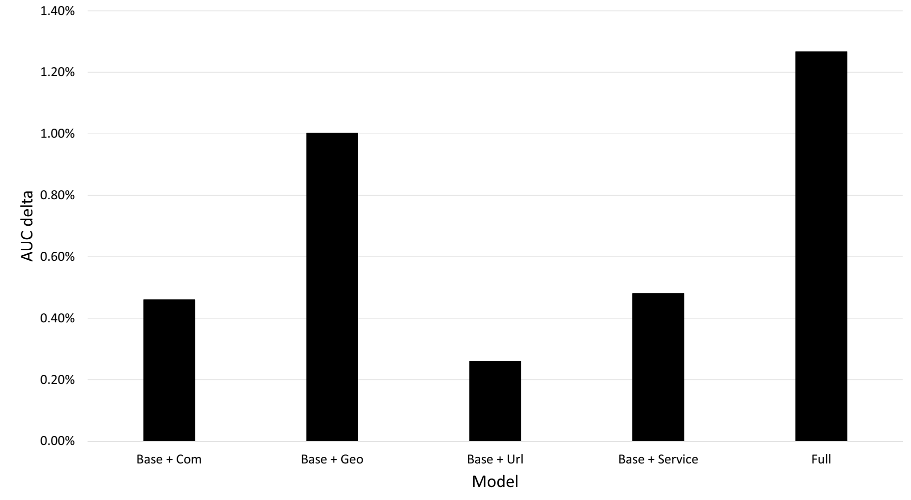

Fig. 1 Incremental AUC value by fine-grained data source for RF (500 trees)

metrics to a full model model including all metrics. Intermediate models include base metric along with one of the 4 fine-grained data sources to investigate this specific data source. The comparison between the results of both benchmark methods namely RF and ERT highlights that RF outperforms ERT for every model based on the same metrics. Consequently, for the discussion of the results below, RF will represent the best benchmark performance. These results are discussed separately according to the contribution.

The added value of multiple fine-grained data sources

The following subsection investigates the added value of multiple fine-grained data sources with the current state of the art i.e. Random Forest. In this context, models with the largest number of trees (i.e. 500 trees) are discussed. Now, let's investigate RQ1, namely the potential to substitute a communication network for churn detection. Using socio-demographical variables and activity variables, the base model reaches an AUC performance of 61.29%. When comparing the additional contribution of features derived from fine-grained data sources in the model, remarkably, the individual incremental value of geo-spatial data (1.00%) largely outperforms the individual contribution of other sources including communication network (0.46%). See Fig. 1.

Concerning RQ2, i.e. the potential to complement the communication network, the full model including all metric categories reaches 62.56% which is 1.27% better than the base model. Consequently, the fine-grained data sources are not fully redundant as their combined performance is higher than any individual ones. This result supports the opportunity to improve churn detection beyond the use of a communication network by means of these additional fine-grained data.

The added value of Essence Random Forest

To better handle the redundancy among features, we propose an adaptation of Random Forest namely Essence Random Forest. Figure 2 compares for these two techniques the incremental AUC among fine-grained data sources. For each types of model, ERF delivers

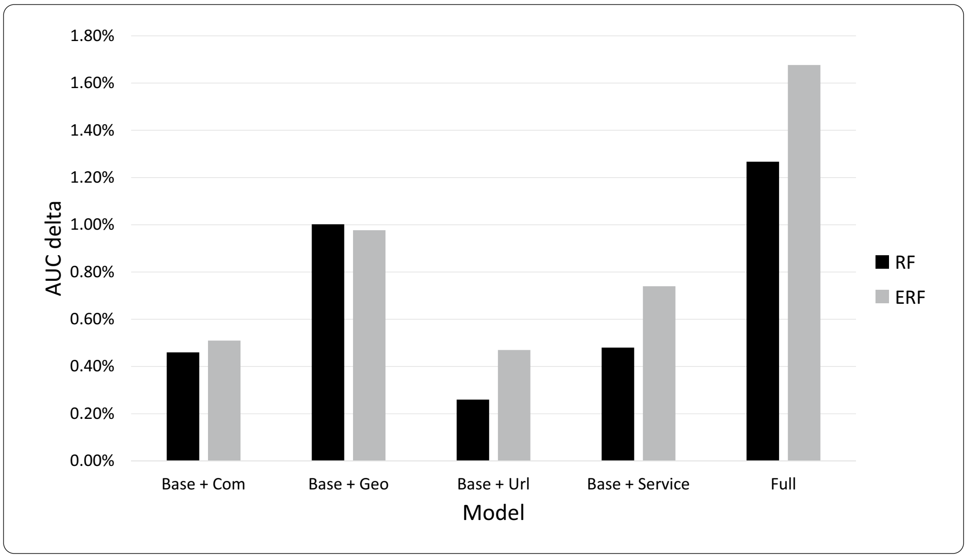

Fig. 2 Incremental AUC value by fine-grained data source: RF VS ERF (500 trees)

a better result or almost the same result for a number of 500 trees. This result answers RQ3 by showing a better classification performance of ERF. This is particularly the case for url, service and the full model while this is less the case for the communication and geo-spatial models. Remarkably, models which benefit the most are the ones which are more likely to contain redundancy by featurization design (as discussed in "Methodology" section). Consistently, the full model which contains all variables is the one which benefit the most from ERF.

What is more, considering multiple levels of trees, the rate of convergence is faster for EFR. This answers to RQ4. In particular, with 10 trees, RF is 1.36% lower than its best result obtained with 500 trees, while ERF is 0.56% lower than the same corresponding results. The comparison between both methods is visually displayed in Fig. 3 for base and full models. While Random Forest and Essence Random Forest equally perform for conventional data, Essence Random Forest outperforms Random Forest when adding fine-grained data not only in classification performance but also in convergence rate. Notably , the best result obtained for RF (corresponding to a number of 200 trees) is only 0.22% better than the result of ERF with 10 trees.

The comparison of Random Forest and Essence Random Forest in terms of variable importance

In order to further investigate the performance difference between RF and ERF, we analyse the variable importance of the underlying models. Figure 4 displays the Random Forest variable importance of the corresponding full model. The most prominent feature i.e. time from enrolment is a base variable but globally, fine-grained data sources features contribute more to this model. This is shown in Fig. 5 where fine-grained data metrics contribute for 76.34% of the model. Among these metrics, service and url metrics contribute the most to this model while communication and geo-spatial metrics contribute

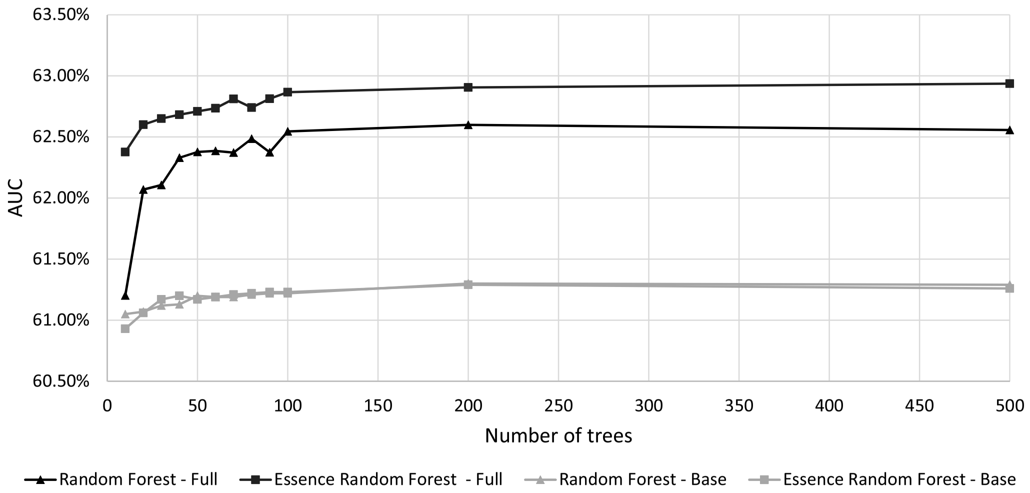

Fig. 3 Comparison AUC performance between RF and ERF for base and full models

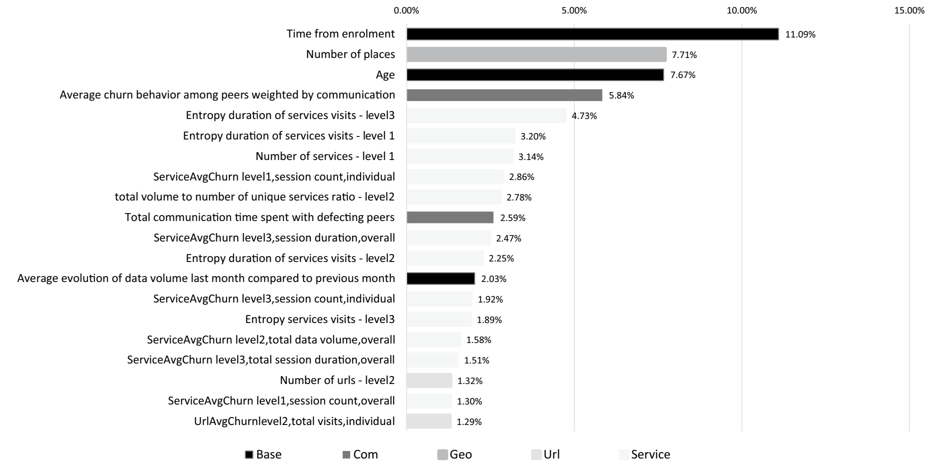

Fig. 4 Random Forest full model (500 trees): variable importance

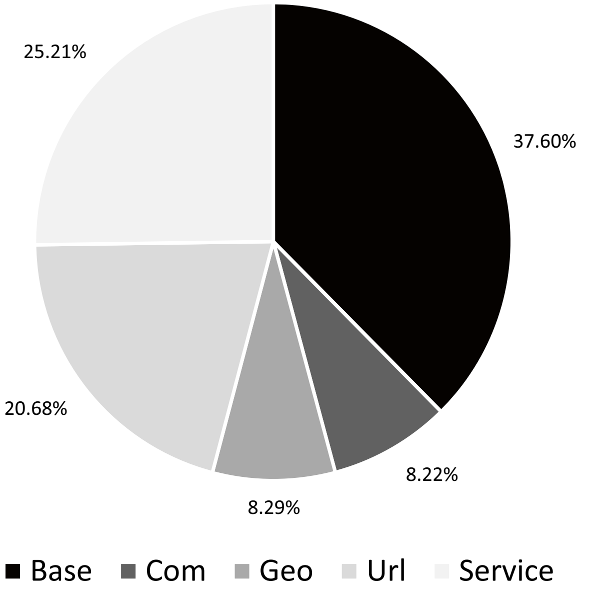

Fig. 5 Variable importance aggregated by metric categories for RF full model (500 trees)

less. Note that these reported aggregated importances correlate to some extent to the number of features derived from the fine-grained data source.

When comparing this result with the Essence Random Forest variable importance of the full model (see Fig. 6), the most important feature remains the same i.e. 'Time from enrolment'. What is however striking is the fact that this feature gets an importance of 24.08%, while it is only 11.09% with RF. Figure 7 displays the same variable importance by metric categories for ERF. The overall importance for base metrics and geo metrics are much higher with ERF than with RF, respectively + 13.94% and + 5.79%. In contrast, this importance is lower for url and service metrics. As discussed above, url metrics and service metrics are more likely to contain redundancy by design than base and

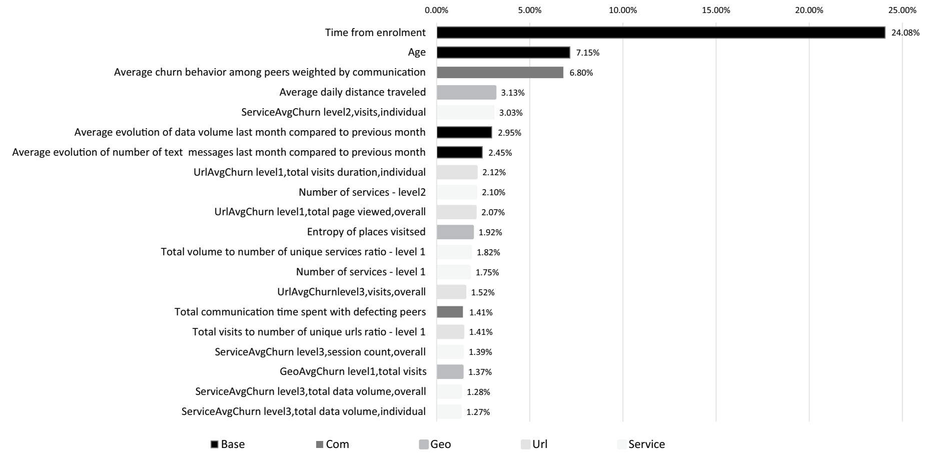

Fig. 6 Essence Random Forest full model (500 trees): variable importance

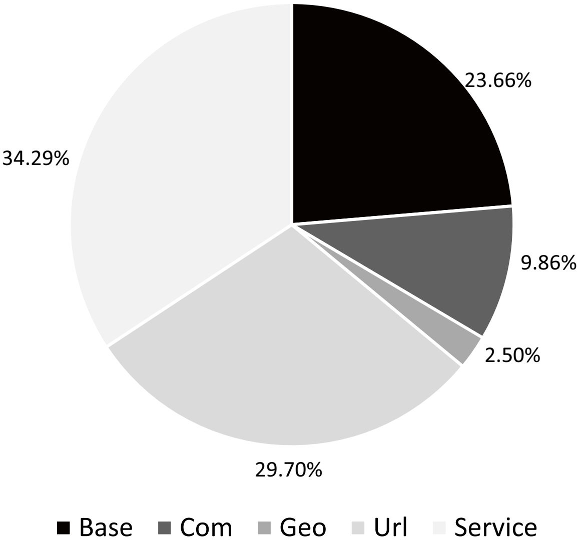

Fig. 7 Variable importance aggregated by metric categories for ERF full model (500 trees)

geo metrics. Consequently, this result suggests that Essence Random Forest better handles redundancy than the original algorithm namely Random Forest. While RF is overwhelmed by the variations of features derived from fine-grained data, ERF rescales the individual weight of each feature to take into account to which extent the information it contains is unique or shared with correlated features.

Discussion

These results confirm the value of alternative internal fine-grained data i.e. websurfing, use of applications and geospatial mobility for churn detection within telephone companies. On the one hand, these data might substitute the widely used communication network. In particular, geo-spatial data is the source leading to the best incremental value. Looking deeper into variable importance, the most contributing features of this category are derived from personality proxies such as the number of visited places. The mobility pattern of the customer is consequently indicative of his/her likelihood of churning. In the current context where classical communications operated by telephone companies are more and more in competition with digital communications operated by actors such as Messenger or Whats'app [8–10], finding alternatives to a communication network likely to disappear might be a critical issue for telephone companies on the long run to keep identifying churn. Our previous work [11] showed that a referral network might constitute a good alternative to this network. This network however requires to build up and manage a referral program, which might be expensive. The current study shows that telephone companies might instead substitute the communication network by leveraging alternative fine-grained data already collected in their daily activities.

On the other hand, investigating the value of these internal fine-grained data on top of a communication network leads to slightly higher incremental value for churn detection. This is a remarkable result as to the best of our knowledge no other study found a complementarity between any fine-grained data and a communication network. In particular, our previous work investigating the value of two context independent networks namely spatial network and referral network [11] found no opportunity to complement a communication network by means of these two networks. This issue is particularly salient as churn detection for telephone companies is critical to survive in saturated markets and in very competitive environments. The cost of retaining a customer is typically assessed to be 5 times less expensive than attracting a new customer [58]. In this context, with a large customer base, even a slightly better identification of likely churners might lead to substantial benefits. This is even more the case that a better identification not only allows to identify more likely churners but it also reduces the number of (false positive) non-churners benefiting from the incentives provided by retention campaigns.

Whatever the perspective investigated, the value of these fine-grained data is larger by using our adaptation of the Random Forest method namely Essence Random Forest. This new method is designed to better handle the redundancy within the data. For fine-grained data, there is usually no standard way to featurise the information that it contains. This leads to multiple feature extraction variations which might be redundant. Our empirical results suggest that this is effectively the case.

In particular, ERF exhibits an higher classification performance. Looking further to the importance of features to build the model, we find that ERF gives more weight to data sources less represented in terms of number of extracted features. More concretely, features coming from tabular data such as socio-demographical information might be overwhelmed by a larger set of features coming from fine-grained data with Random Forest. From a managerial perspective, the use of ERF algorithm would consequently allow to better leverage these fine-grained data either from a substitution perspective or a complementarity perspective discussed above.

What is more, ERF faster converges to a stable result. This might also lead to supplementary monetary benefits on a IT infrastructure where the cost heavily depends on the use. In cloud solutions for example, the cost of computation is typically more expensive than the cost of storage. Based on our results, using ERF with 10 trees leads to the same classification performance of RF with 50 trees. In this scenario, the cost of computation is consequently 80% less expensive.

From a theoretical perspective, the classification performance of Random Forest is driven by the performance of individual trees and the diversity of trees [17]. Notably, Extremely Randomized Trees fosters a lower correlation between trees by giving random thresholds to features but it might lead to a lower classification performance of individual trees. By its mechanism of giving more weight to information less represented in the feature set, Essence Random Forest works on both sources of classification performance to improve Random Forest in the presence of redundancy.

Conclusion

The exponential growth of fine-grained data leads to major promises in the domain of marketing analytics as long as new methods are developed to leverage it [2]. In particular, the appropriate use of fine-grained data for churn detection within telephone companies receives a special interest as this line of research has received few attention from the literature so far, beyond the use of a communication network [7]. To the best of our knowledge, only the work of [11] highlights the value of an alternative network namely a referral network. This network however requires to build up and manage a referral program which might be expensive.

In the present study, we investigate the value of alternative fine-grained data for churn detection which are already collected by telephone companies in their daily operations. This value is assessed from two perspectives: from a substitution perspective on the long term and from a complementarity perspective on the short term. Furthermore, we also investigate how to handle it.

The featurization process of fine-grained data is more likely to deliver multiple variations of the same information leading to more correlated features which might prevent current methods to optimally handle it. Therefore, we propose a new version of the seminal Random Forest method called Essence Random Forest. This method is designed to be more robust in the presence of redundant information. In particular, we investigate if this variation offers a better classification performance and a better convergence to stable results.

To assess these two contributions, data from an European telephone company is analysed in a churn detection modelling context. In particular, we propose multiple feature extractions ariations of raw communication, mobility, url and service data. Our results show that metrics derived from geo-spatial mobility outperform metrics from the well-known communication network. Consequently, geo-spatial metrics might become a good alternative to communication network metrics as classical communication might disappear in the future in favour of digital communications [8–10]. Furthermore, the Random Forest full model slightly gets an higher performance than any individual model with one category of features which indicates that information contained in these data are not fully redundant. To the best of our knowledge, this article is the first to highlight the value of alternative networks on top of a well-known communication network for churn detection. For managers of telephone companies, this insight is of utter importance to tackle churn behaviour in a saturated market.

These two sources of value for churn prediction i.e. substitution perspective and complementarity perspective are better leveraged with Essence Random Forest. Indeed, while Random Forest and Essence Random Forest equally perform on conventional data, we find that Essence Random Forest offers a better classification performance when using these fine-grained data. In addition, Essence Random Forest also converges faster to a stable result. These better results may be explained by the fact that Essence Random Forest rescales the probability of each feature of being selected to split a node depending on the uniqueness of the embedded information it contains. These two advantages i.e. classification and convergence might both contribute to supplementary monetary benefits. First, the benefits of the substitution perspective and complementarity perspective discussed above are enhanced to an even higher level. Second, in a IT environment with a substantial usage cost, ERF already provides a close to state of the art classification performance at a fraction of the current cost. From a theoretical perspective, ERF, by giving more weights to information less represented in the feature set, fosters a better classification performance and an higher diversity of trees. These two effects combined affects positively the performance of the resulting model.

In terms of limitations, the empirical results are based on the data from one telephone company. In order to ensure the generalisation capability of the results to other telephone companies, it would be interesting to analyse also the data of other telephone companies. From a practical point of view, this would be however a tricky task for two main reasons. First, accessing data of a telephone company is subject to thorough discussions with telephone companies. Second, mobile data generates a high volume of data, their treatment requires an appropriate infrastructure which is already demanding with the data of one telephone company. For instance, roughly 20 Terabytes of raw data were analysed for this study. This is why empirical studies related to mobile data typically focus on the data of a single telephone company. However, we believe that the generalisation capability of the results is supported by the fact that telephone companies typically collect the same type of data.

Future studies might investigate the value of Essence Random Forest in other data modelling contexts such as new fine-grained data or other application domains to further support its relevance to handle redundant data. In particular, genomic data analysis seems a promising field of investigation [59]. Indeed, Random Forest is a state of the art technique in this specific context. Furthermore, genomic data is characterized by a large number of features which largely exceeds the number of observations leading to an high level of redundancy.

More generally, this study introduces a new key idea behind Essence Random Forest to reduce redundancy. It points to the fact that simple random selection of features favours information supported by redundant features. This issue will become more and more pressing in the current context of a ever growing redundancy among features due to the explosion of unstructured data. This idea could be useful to also improve other machine learning techniques as well. For example, the dropout mechanism used for deep learning might deserve some attention in this context as it consists of randomly dropping nodes in the deep neural network [60]. Consequently, assigning a different probability of dropout according to the uniqueness of the information given by the node might lead to improve the results of this mechanism.

Acknowledgements
Not applicable.

Authors' contributions
All authors conceived and designed the analysis. CC and PB collected the data. CC proposed a new algorithm (i.e. Essence Random Forest) and performed the analysis. All authors contributed to the writing of the paper. All authors read and approved the final manuscript.

Authors' information
Christian Colot is a Postdoctoral Researcher at the University of Namur in Belgium, Department of Business Administration. He has recently finished his PhD in Management sciences. He also holds a Master degree in Information System from the University of Lille and Masters degrees in Statistics and Management sciences from UCLouvain. Christian is member of the Namur Digital Institute. His research interests include Customer Relationship Management, Big Data Analytics and Mobile data. His work has appeared in several international peer reviewed conferences and journals.

Philippe Baecke is an Associate Professor at Vlerick Business School (Belgium) and Adjunct Professor at Trinity Business School (Ireland) with a strong expertise in big data analytics. He is programme director of the MSc in "Digital & Marketing Management" at Vlerick Business School and two executive programmes: "Creating Business Value with Big Data" & "Data Driven Marketing". From a research perspective, Philippe focuses on improving business insights by creatively incorporating new data types, such as geographical and social network data. His research has been published in several peer reviewed journals such as International Journal of Operations and Production Management, Omega: the International Journal of Management Science, Decision Support Systems, International Journal of Productions Economics and others.

Isabelle Linden is a Professor of Information Management at the University of Namur in Belgium, Department of Business Administration. She obtained her PhD in Computer Sciences from the University of Namur. She also holds Masters degrees in Philosophy and in Mathematics from the University of Liège, Belgium. She is member of the FoCuS Research Group of the Namur Digital Institute. Combining theoretical computer science and business administration, her main research domain regards information, knowledge and artificial intelligence. She explores their integration within systems as EIS, DSS and BI systems. Her works can be found in several international edited books, journals, books chapters and conferences. She serves as reviewer and program committee member in several international journals, conferences and workshops.

Funding
Not applicable.

Data availability statement
The dataset analysed during the current study is not publicly available due to a Non Disclosure Agreement with the European telephone company which gave access to this data.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Business Administration, University of Namur, Namur, Belgium. [2]Vlerick Business School, Gent, Belgium.

Received: 13 January 2021 Accepted: 13 April 2021
Published online: 29 April 2021

References

1. Coughlin T. 175 Zettabytes By 2025. https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/
2. Wedel M, Kannan P. Marketing analytics for data-rich environments. J Market. 2016;80(6):97–121.
3. Perlich C, Dalessandro B, Raeder T, Stitelman O, Provost F. Machine learning for targeted display advertising: transfer learning in action. Mach Learn. 2014;95(1):103–27.
4. Al-Zuabi IM, Jafar A, Aljoumaa K. Predicting customer's gender and age depending on mobile phone data. J Big Data. 2019;6(1):18.
5. Lismont J, Ram S, Vanthienen J, Lemahieu W, Baesens B. Predicting interpurchase time in a retail environment using customer-product networks: an empirical study and evaluation. Exp Syst Appl. 2018;104:22–32.
6. Martens D, Provost F, Clark J, de Fortuny EJ. Mining massive fine-grained behavior data to improve predictive analytics. MIS Q. 2016;40:4.
7. Ascarza E, Neslin SA, Netzer O, Anderson Z, Fader PS, Gupta S, Hardie BG, Lemmens A, Libai B, Neal D, et al. In pursuit of enhanced customer retention management: review, key issues, and future directions. Customer Needs Sol. 2018;5(1–2):65–81.

8. Farooq M, Raju V. Impact of over-the-top (OTT) services on the telecom companies in the era of transformative marketing. Global J Flexible Syst Manag. 2019;20(2):177–88.
9. Stork C, Esselaar S, Chair C. OTT-Threat or opportunity for African Telcos? Telecommun Policy. 2017;41(7–8):600–16.
10. Sujata J, Sohag S, Tanu D, Chintan D, Shubham P, Sumit G. Impact of over the top (OTT) services on telecom service providers. Indian J Sci Technol. 2015;8(S4):145–60.
11. Colot C, Baecke P, Linden I. Alternatives for Telco Data Network: the value of spatial and referral networks for churn detection. Inf Syst Manag. 2021;8:1–19.
12. Lessmann S, Baesens B, Seow H, Thomas L. Benchmarking state-of-the-art classification algorithms for credit scoring: a 10-year update. Eur J Operat Res. 2015;247(1):124–36.
13. Louppe G. Understanding random forests: From theory to practice. 2014; arXiv preprint arXiv:1407.7502
14. Ngai EW, Xiu L, Chau DC. Application of data mining techniques in customer relationship management: a literature review and classification. Exp Syst Appl. 2009;36(2):2592–602.
15. Ahmad AK, Jafar A, Aljoumaa K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data. 2019;6(1):28–51.
16. Al-Molhem NR, Rahal Y, Dakkak M. Social network analysis in telecom data. J Big Data. 2019;6(1):99.
17. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
18. Gao Y-F, Li B-Q, Cai Y-D, Feng K-Y, Li Z-D, Jiang Y. Prediction of active sites of enzymes by maximum relevance minimum redundancy (mrmr) feature selection. Mol BioSyst. 2013;9(1):61–9.
19. Idris A, Rizwan M, Khan A. Churn prediction in telecom using random forest and PSO based data balancing in combination with various feature selection strategies. Comput Elect Eng. 2012;38(6):1808–19.
20. Kandaswamy KK, Pugalenthi G, Kalies K-U, Hartmann E, Martinetz T. Ecmpred: Prediction of extracellular matrix proteins based on random forest with maximum relevance minimum redundancy feature selection. J Theor Biol. 2013;317:377–83.
21. Li B-Q, Feng K-Y, Chen L, Huang T, Cai Y-D. Prediction of protein-protein interaction sites by random forest algorithm with MRMR and IFS. PloS ONE. 2012;7(8):43927.
22. Liu L, Chen L, Zhang Y-H, Wei L, Cheng S, Kong X, Zheng M, Huang T, Cai Y-D. Analysis and prediction of drug-drug interaction by minimum redundancy maximum relevance and incremental feature selection. J Biomol Struct Dyn. 2017;35(2):312–29.
23. Ma X, Sun X. Sequence-based predictor of atp-binding residues using random forest and MRMR-IFS feature selection. J Theor Biol. 2014;360:59–66.
24. Ma X, Guo J, Sun X. Sequence-based prediction of RNA-binding proteins using random forest with minimum redundancy maximum relevance feature selection. BioMed Res Int. 2015;2015:78.
25. Jan ZM, Verma B. Ensemble classifier optimization by reducing input features and base classifiers. In: 2019 IEEE congress on evolutionary computation (CEC). IEEE, 2019;1580–1587.
26. Amaratunga D, Cabrera J, Lee Y-S. Enriched random forests. Bioinformatics. 2008;24(18):2010–4.
27. Nagpal A, Singh V. Identification of significant features using random forest for high dimensional microarray data. J Eng Sci Technol. 2018;13(8):2446–63.
28. Nguyen T-T, Huang JZ, Nguyen TT. Unbiased feature selection in learning random forests for high-dimensional data. Sci World J. 2015;2015:7.
29. Wang Q, Nguyen T-T, Huang JZ, Nguyen TT. An efficient random forests algorithm for high dimensional data classification. Adv Data Anal Classif. 2018;12(4):953–72.
30. Wu Q, Ye Y, Liu Y, Ng MK. SNP selection and classification of genome-wide SNP data using stratified sampling random forests. IEEE Trans Nanobiosci. 2012;11(3):216–27.
31. Xu B, Huang JZ, Williams G, Wang Q, Ye Y. Classifying very high-dimensional data with random forests built from small subspaces. IJDWM. 2012;8(2):44–63.
32. Xu B, Huang JZ, Williams G, Ye Y. Hybrid weighted random forests for classifying very high-dimensional data. Int J Data Warehous Mining. 2012;8(2):44–63.
33. Ye Y, Wu Q, Huang JZ, Ng MK, Li X. Stratified sampling for feature subspace selection in random forests for high dimensional data. Pattern Recogn. 2013;46(3):769–87.
34. Zhang Y, Cao G, Li X, Wang B. Cascaded random forest for hyperspectral image classification. In: IEEE journal of selected topics in applied earth observations and remote sensing. 2018;11(4):1082–94.
35. Kyrillidis A, Zouzias A. Non-uniform feature sampling for decision tree ensembles. In: 2014 IEEE international conference on acoustics, speech and signal processing. IEEE, 2014;4548–4552.
36. Rodriguez JJ, Kuncheva LI, Alonso CJ. Rotation forest: a new classifier ensemble method. IEEE Trans Pattern Anal Mach Intellig. 2006;28(10):1619–30.
37. Zhang L, Suganthan PN. Random forests with ensemble of feature spaces. Pattern Recogn. 2014;47(10):3429–37.
38. Carreira-Perpiñán MÁ, Zharmagambetov A. Ensembles of Bagged TAO Trees Consistently Improve over Random Forests, AdaBoost and Gradient Boosting. In: Proceedings of the 2020 ACM-IMS on foundations of data science conference, 2020; p 35–46
39. Katuwal R, Suganthan PN, Zhang L. Heterogeneous oblique random forest. Pattern Recognition. 2020;99.
40. Rastogi R, David A. Oblique Random Forest via Regularized Multisurface Proximal Support Vector Machine. In: 2019 Global conference for advancement in technology (GCAT). IEEE, 2019; p 1–6.
41. Zhang L, Varadarajan J, Nagaratnam Suganthan P, Ahuja N, Moulin P. textbfRobust visual tracking using oblique random forests. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017; p 5589–98
42. Abellan J, Mantas CJ, Castellano JG, Moral-Garcia S. Increasing diversity in random forest learning algorithm via imprecise probabilities. Exp Syst Appl. 2018;97:228–43.
43. Mantas CJ, Castellano JG, Moral-García S, Abellán J. A comparison of random forest based algorithms: random credal random forest versus oblique random forest. Soft Comput. 2019;23(21):10739–54.
44. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3–42.
45. Sun J, Zhong G, Dong J, Saeeda H, Zhang Q. Cooperative profit random forests with application in ocean front recognition. IEEE Access. 2017;5:1398–408.

46. Sun J, Zhong G, Huang K, Dong J. Banzhaf random forests: cooperative game theory based random forests with consistency. Neural Netw. 2018;106:20–9.
47. Zhang Y, Song B, Zhang Y, Chen S. An Advanced Random Forest Algorithm Targeting the Big Data with Redundant Features. In: International conference on algorithms and architectures for parallel processing. Springer, 2017; p 642–51.
48. Bernstein MN. Note on Random Forests. http://pages.cs.wisc.edu/~lowmatthewb/pages/notes/pdf/ensembles/RandomForests.pdf
49. SAS Institute Inc.: SAS/STAT 15.1 User's Guide, 2018;
50. Ma L, Krishnan R, Montgomery AL. Latent homophily or social influence? an empirical analysis of purchase within a social network. Manag Sci. 2014;61(2):454–73.
51. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: homophily in social networks. Ann Rev Sociol. 2001;27:415–44.
52. Cialdini RB, Goldstein NJ. Social influence: compliance and conformity. Annu Rev Psychol. 2004;55:591–621.
53. Stankova M, Martens D, Provost F. Classification over bipartite graphs through projection 2015.
54. De Montjoye Y-., Quoidbach J, Robic F, Pentland A. Predicting Personality Using Novel Mobile Phone-based Metrics. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 7812 LNCS, 2013;48–55. Cited By :159. www.scopus.com
55. McCrae RR, John OP. An introduction to the five-factor model and its applications. J Personal. 1992;60(2):175–215.
56. Breiman L, Last M, Rice J. Random forests: finding quasars. Statistical challenges in astronomy. New York: Springer; 2003. p. 243–54.
57. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;22:837–45.
58. Wertz J. Don't Spend 5 Times More Attracting New Customers, Nurture The Existing Ones. 2018. https://www.forbes.com/sites/jiawertz/2018/09/12/dont-spend-5-times-more-attracting-new-customers-nurture-the-existing-ones/?sh=4a8dd4b25a8e
59. Zaim SR, Kenost C, Berghout J, Chiu W, Wilson L, Zhang HH, Lussier YA. binomialRF: interpretable combinatoric efficiency of random forests to identify biomarker interactions. BMC Bioinf. 2020;21(1):1–22.
60. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15(1):1929–58.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.