Predicting and Explaining Privacy Risk Exposure in Mobility Data

Francesca Naretto[1], Roberto Pellungrini[2](✉), Anna Monreale[2], Franco Maria Nardini[3], and Mirco Musolesi[4,5]

[1] Scuola Normale Superiore, Pisa, Italy
francesca.naretto@sns.it
[2] University of Pisa, Pisa, Italy
{roberto.pellungrini,anna.monreale}@di.unipi.it
[3] ISTI CNR, Pisa, Italy
francomaria.nardini@isti.cnr.it
[4] University College London, London, UK
m.musolesi@ucl.ac.uk
[5] University of Bologna, Bologna, Italy

Abstract. Mobility data is a proxy of different social dynamics and its analysis enables a wide range of user services. Unfortunately, mobility data are very sensitive because the sharing of people's whereabouts may arise serious privacy concerns. Existing frameworks for privacy risk assessment provide tools to identify and measure privacy risks, but they often (i) have high computational complexity; and (ii) are not able to provide users with a justification of the reported risks. In this paper, we propose EXPERT, a new framework for the prediction and explanation of privacy risk on mobility data. We empirically evaluate privacy risk on real data, simulating a privacy attack with a state-of-the-art privacy risk assessment framework. We then extract individual mobility profiles from the data for predicting their risk. We compare the performance of several machine learning algorithms in order to identify the best approach for our task. Finally, we show how it is possible to explain privacy risk prediction on real data, using two algorithms: SHAP, a feature importance-based method and LORE, a rule-based method. Overall, EXPERT is able to provide a user with the privacy risk and an explanation of the risk itself. The experiments show excellent performance for the prediction task.

Keywords: Privacy risk assessment · Privacy risk prediction · Explainability

1 Introduction

There is a growing research interest in mobility data analysis, since it is a key enabler of a new wave of knowledge-based services and applications. However, the use of human mobility data raises concerns associated to the potential leakage of personal sensitive information as mobility data analysis might reveal details

A. Appice et al. (Eds.): DS 2020, LNAI 12323, pp. 403–418, 2020.
https://doi.org/10.1007/978-3-030-61527-7_27

of people's private life. For example, de Montjoye *et al.* [17] showed that four spatio-temporal points can be enough to uniquely identify 95% of the individuals in a mobility dataset. The existence of these privacy issues has led researchers to develop techniques to mitigate the privacy risks while preserving mobility data [4,15,26]. For enabling a practical application of these techniques, Pratesi *et al.* proposed PRUDEnce [21], a framework for a systematic assessment of individual privacy risk in a mobility dataset. PRUDEnce helps data controllers being compliant with the new EU General Data Protection Regulation (GDPR)[1]. However, PRUDEnce is characterized by a high computational complexity, because it requires the computation of the maximum risk of re-identification (or privacy risk) given an external knowledge that a malicious adversary might use for an attack [20]. The high computational complexity becomes a non-negligible practical limitation in some online user-centric applications where it is useful to have a continuously up-to-date indicator of privacy exposure. In user-centric applications, providing users with an explanation of the reasons of the identified privacy risk might contribute to raise their self-awareness.

In this paper, to overcome the computational complexity drawback and to increase users' awareness, we propose EXPERT, an EXplainable Privacy ExposuRe predicTion framework that exploits *(i)* machine learning (ML) models for predicting a user's individual privacy risk and *(ii) local* explainers for producing explanations of the predicted risk. First, EXPERT extracts from human mobility data an individual mobility profile describing the mobility behavior of any user. Second, for each user it exploits PRUDEnce to compute the associated privacy risk. Third, it uses the mobility profiles of the users with their associated privacy risks to train a ML model. For the prediction task, EXPERT exploits tree-based ensemble models to effectively handle the class-imbalance problem, i.e., a high number of risky users vs a low number of non-risky ones, that is typical of the data in this context. The aim is to have a predictor that preserves the privacy of risky users while providing the freedom of using data-driven services to users with low privacy risk. For a new user, along with the prediction of risk, EXPERT also provides an explanation of the predicted risk. EXPERT exploits two state-of-the-art explanation techniques, i.e., SHAP [13] and LORE [11]. The two methods produce explanations based on feature importance and logic rules, respectively. The goal of explanations is to provide users with insights on which mobility behavior contributes to their privacy risk. We evaluate EXPERT on real-world mobility data showing the effectiveness of the framework. Results show that the proposed framework is able to classify the privacy risk level of unseen users in the urban areas. Moreover, we observe a high recall on the high-risk users, meaning that the probability of misclassifying a high-risk user as low-risk is negligible, while achieving good performance in classifying low-risk users.

The paper is organized as follows. Sect. 2 discusses related work. In Sect. 3, we briefly discuss PRUDEnce, the framework we used for the privacy risk assessment. Section 4 introduces our novel EXPERT framework. In Sect. 5, we report

[1] EU GDPR can be found at the following link: http://bit.ly/1TlgbjI.

the results of a comprehensive experimental evaluation of EXPERT on mobility data. Finally, Sect. 6 concludes the work and discusses future work.

2 Related Work

Our framework leverages the privacy risk assessment framework PRUDEnce [21], which allows for the systematic calculation of the empirical privacy risk. Another risk management framework is LINDDUN [8] useful for modeling privacy threats in software-based systems, but lacks a quantitative evaluation of privacy risk. Some works [23,25] propose to evaluate the privacy risk by a unicity measure computed as the number of records uniquely identified. Armando *et al.* [2] proposed a risk-aware framework for information disclosure supporting runtime risk assessment where access-control decisions are based on the disclosure-risk associated with a data access request and adaptive anonymization is used as a risk-mitigation method.

In the context of mobility analysis, an overview on problems, techniques and methodologies can be found in [28]. Human mobility analysis can reveal personal sensitive information and habits leading to possible privacy violation. Thus, many techniques for privacy-preserving analysis have shown that we can design data-driven mobility services where the quality of results coexists with the privacy protection. Some works, e.g., [4,16], are based on the differential privacy model [9] while others, e.g., [15,26], are based on the k-anonymity model [24].

Our work can be seen as an extension of the prediction methodology proposed by Pellungrini *et al.* [20], showing how it is possible to predict privacy risk in mobility data with a feature based approach. We extend it by providing a unified framework that provides both prediction and explanation about the individual privacy risk. Moreover, our proposal is based on a prediction module that is able to handle the high class imbalance of the data typical of this domain [29].

The importance of interpretability in machine learning has led to an increasing research work in this field. An overview of explainable machine learning models can be found in [12]. This survey identifies two main families of approaches: *local* and *global* explainers. The first category aims at explaining the reason for a specific instance classification [11,13,22], while the goal of the second one is to explain the logic of the "machine learning black-box" as a whole [5–7].

3 Background

Human mobility data contain information about the movement of individuals during a given period of observation. They are typically collected by electronic devices, such as mobile phones and GPS devices installed in vehicles [28]. All the movements of a user in the period of observation are described using a sequence of spatio-temporal data points, i.e., a trajectory. In other words, each sequence item is a pair composed of a geographic location, often expressed in coordinates (generally latitude and longitude), and a timestamp indicating when the user stopped in or went through that location.

Definition 1 (Trajectory). *A human mobility trajectory is a temporally ordered sequence of pairs,* $T_u = (l_1, t_1), (l_2, t_2), \ldots, (l_m, t_m)$*, where* $l_i = \langle x_i, y_i \rangle$ *is the location identified by the latitude* x_i *and longitude* y_i*, while* t_i *(*$i = 1, \ldots, m$*) denotes the corresponding timestamp such that* $\forall 1 \leq i \leq m \; t_i < t_{i+1}$*.*

We denote by $\mathcal{D} = T_1, \ldots, T_n$ the *mobility dataset* that describes the complete history of movements of n individuals, in a specific period of observation.

3.1 Privacy Risk Assessment Framework

In this paper, we consider the framework PRUDEnce [21], which allows for a systematic assessment of the privacy risk inherent to human mobility data. It considers a scenario where a Service Developer (SD) requests data from a Data Provider (DP) to develop services or perform an analysis. In order to guarantee the right to privacy of individuals, the DP has to assess their privacy risk before the data sharing. Once assessed the privacy risk, the DP can choose how to protect the data before sharing them, selecting the most appropriate privacy-preserving technology. Taking into account the data requirements of the SD, the DP aggregates, selects, and filters the dataset $\mathcal{D}$ to meet its requirements and on top of it performs a privacy risk assessment. This operation requires the definition of a set of possible attacks that an adversary might conduct on the data, and their simulation. The user's privacy risk is related to her probability of re-identification in a dataset with respect to a set of attacks. An attack assumes that an adversary gets access to a dataset, then, using some previously obtained background knowledge, i.e., the knowledge of a portion of an individual's mobility data, the adversary tries to re-identify all the records in the dataset regarding that individual. An attack is defined by a matching function, which represents the process with which an adversary exploits the background knowledge to find the corresponding individual in the data. As far as the attack definition is concerned, PRUDEnce is based on the notions of background knowledge category, configuration and instance. The first one denotes the type of information known by the adversary about a specific set of dimensions of an individual's mobility data: e.g.., a subset of the locations visited by a user (spatial dimension) or the specific times a user visited those locations (spatial and temporal dimensions). The number of the elements known by the adversary is called background knowledge configuration. An example is the adversary knowledge of $h = 2$ locations visited by an individual. Finally, an instance of background knowledge is defined as the specific information known by the adversary, such as a visit in a specific location. Consider a trajectory from D: $T_u = \langle (l_1, t_1), (l_2, t_2), (l_3, t_3), (l_4, t_4) \rangle$ of an individual u. Based on T_u the DP can generate all the possible instances of a background knowledge configuration that an adversary might use to re-identify the whole T_u. If the adversary knows the ordered subsequences of locations and $h = 2$, we obtain the background knowledge configuration: $B_2 = \{((l_1, t_1), (l_2, t_2)),\ ((l_1, t_1), (l_3, t_3)),\ ((l_1, t_1), (l_4, t_4)),\ ((l_2, t_2), (l_3, t_3)),\ ((l_2, t_2), (l_4, t_4)),\ ((l_3, t_3), (l_4, t_4))\}$. The adversary might know instance $b = ((l_1, t_1), (l_4, t_4)) \in B_{h=2}$ and aims at detecting all the records in D regarding u, in order to reconstruct the whole trajectory T_u.

The definition of privacy risk is based on these notions and on the following definition of probability of re-identification.

Definition 2. *Given an attack and its function* $matching(T, b)$ *indicating if a record* $T \in \mathcal{D}$ *matches the instance of background knowledge configuration* $b \in B_h$*, and a function* $M(\mathcal{D}, b) = \{T \in \mathcal{D} | matching(T, b) = True\}$*, we define the probability of re-identification of an individual* u *in dataset* $\mathcal{D}$ *as:* $PR_{\mathcal{D}}(T = u|b) = \frac{1}{|M(\mathcal{D},b)|}$ *that is the probability to associate a record* $T \in \mathcal{D}$ *to an individual* u*, given instance* $b \in B_h$*.*

Since each instance $b \in B_h$ has its own probability of re-identification, the risk of re-identification of an individual is defined as the maximum probability of re-identification over the set of instances of a background knowledge configuration:

Definition 3. *The risk of re-identification (or privacy risk) of an individual* u *given a background knowledge configuration* B_h *is her maximum probability of re-identification* $Risk(u, \mathcal{D}) = \max PR_{\mathcal{D}}(T = u|b)$ *for each* $b \in B_h$*.*

4 Explainable Privacy Risk Prediction Framework

PRUDEnce [21] assumes a worst case scenario approach for the privacy risk computation and therefore, it evaluates all the possible background knowledge configurations for a potential adversary generating them with a combinatorial approach directly from the data of a user. While the framework provides a comprehensive methodology for worst-case privacy risk assessment, its computational complexity is high. Moreover, PRUDEnce is designed for supporting data providers (companies) in identifying portions of data with high privacy risk by simulations of the attacks. The computation requires the availability of the entire dataset, like that stored in the servers of the companies. In other words, PRUDEnce is not suited for providing personalized recommendations in terms of risks associated to sharing personal trajectories. Indeed, for any new user requiring risk evaluation, the system should re-compute the privacy risk against the whole dataset. Moreover, it does not provide any explanation of the privacy risk derived by the system. In this paper we present an explainable framework for the *individual* prediction of a user's privacy risk, in order to increase privacy risk awareness, by also providing an explanation of the derivation of the risk associated to sharing sensitive location information. The idea is inspired by the explainable privacy-preserving system theorized in [3]. To this end, we propose EXPERT which, given a user's trajectory, predicts the privacy risk associated with it. The explanation provided to the users is based on their trajectory given in input. Figure 1 depicts the architecture of EXPERT which is composed of two main modules: the *privacy risk prediction* module which takes as input the user's trajectory and, exploiting a trained ML model, predicts the privacy risk level of that user, and the *explanation* module which produces the explanation of the predicted risk. The ML model is the result of several steps: *(i)* the empirical computation of the individual privacy risk, *(ii)* the extraction of individual mobility profiles from human mobility data, summarizing users' mobility behavior, and *(iii)* the training of a ML model.

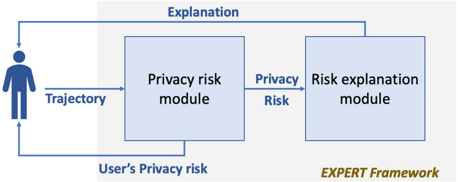


Fig. 1. The general structure of the proposed framework EXPERT.

4.1 Learning a Prediction Model for Individual Privacy Risk

The basic idea is to train a ML model to predict the privacy risk level of users based solely on their individual mobility profile. Thus, given a human mobility dataset of n user trajectories, we propose to derive the training dataset $\langle M, \Gamma\rangle$, where M is a set of n individual mobility profiles, and Γ is the vector of their associated privacy risk levels. Since, the privacy risk is related to a specific attack (see Sect. 3.1), the procedure for *building a training dataset* depends on the adversary attack modelling. As a consequence, given a specific attack, characterized by a background knowledge configuration B_h, the procedure performs the following two steps:

- *Mobility Profile Extraction*: Given a mobility dataset $\mathcal{D}$, for every user trajectory T_u we propose to extract a mobility profile in order to characterize her mobility behavior. To this end, we propose to derive a set of well-known mobility features (presented in the next section). We denote by $M_u \in M$ the mobility feature vector of a user.
- *Privacy Risk Computation*: For each user u a privacy risk value is computed by simulating an attack with background knowledge configuration B_h on the mobility dataset $\mathcal{D}$. Since the goal is to predict the privacy risk level, the privacy risk vector is discretized to get a set of risk classes[2], and the vector of n user's privacy risk levels $\mathit{\Gamma}$.

After the execution of the above two steps, we get a training set $\langle M, \mathit{\Gamma}\rangle$. The derived training dataset $\langle M, \Gamma\rangle$ is used to train a predictive model which will be used within EXPERT to immediately estimate the privacy risk level of previously unseen users, whose data were not used in the learning process. Clearly, in prediction time, in order to predict the privacy risk of a new trajectory instance the process requires, first the computation of the mobility profile for that user and

[2] In our experiments we discretize the risk in two main classes: low risk (privacy risk ≤ 0.5) and high risk (privacy risk > 0.5).

then, the application of the predictive model. Among the different ML methods, we propose to employ models able to handle classification tasks with imbalanced data. Indeed, as we show in our experiments, one of the characteristics of our training data is that most of the users have high privacy risk. Our goal is to get a predictor able to guarantee the privacy protection of risky users while providing the freedom of using data-driven services to users with low privacy risk. Thus, the optimal predictor should be characterized by a low probability of misclassifying a high risk user as a low risk one, while maintaining also good performance with respect to the classification of low risk users. In this paper, we propose to apply the gcForest model [29], a decision tree ensemble approach with performance highly competitive to deep neural networks in a broad range of tasks. It is especially suitable to handle highly extra-imbalanced data [27]. gcForest relies on multiple layers of parallel forests of trees whose output is then concatenated to re-represent data to subsequent layers. In our experiments we compare gcForest against models such as decision tree, logistic regression, and random forest.

Mobility Profile Extraction. The goal of this step is to construct the matrix M representing the set of individual mobility profiles, expressed by a set of mobility features that describe and summarize the mobility behavior of an individual. In our setting, we employ measures widely used in the literature [18,20]. Some of them describe only the mobility behaviour of an individual, while others describe an individual mobility behaviour in relation to collective mobility characteristics. Table 1 reports all the mobility measures used in the study. First of all, we define V as the number of visits of a user, it corresponds to the total number of locations in the user's trajectory. To quantify the erratic behaviour of a user during the day we compute the average number of daily visits $\overline{V}$, dividing V by the total number of days in the period of observation. *Locs*, instead, is the number of distinct locations visited by a user during the period of observation, while $Locs_{ratio}$ represents the fraction of locations covered by a user. We compute it by dividing *Locs* by the total number of locations available in the territory. We also evaluated some measures about the distances travelled by the users. We define D_{max} as the maximum distance travelled by each user, i.e. the longest trip for each user. This measure is then employed for the computation of D^{trip}_{max}: it is the ratio between the maximum distance travelled D_{max} and the maximum distance that is possible to travel in the area of observation. We also consider D_{sum}, i.e., the sum of all the distances travelled by a user. This value is then used in the definition of $\overline{D_{sum}}$, which is the average of D_{sum} over the period of observation (expressed in days). We also consider the *radius of gyration* [19] representing the characteristic distance travelled by a user during the period of observation and is defined as $r_g = \sqrt{\frac{1}{N}\sum_{i \in L} w_i (r_i - r_{cm})^2}$, in which $i \in L$ is the visited location by a user, w_i represents a user's frequency of visits at a location i, r_i denotes the geographical description of the location i and it is a bi-dimensional vector, while r_{cm} is the center of mass of the user under consideration. Mathematically, the latter is defined as $r_{cm} = \frac{1}{V}\sum_{1 \in L} r_i$. We also measure the *mobility entropy* E as the predictability of a user's trajectory. We employ the Shannon entropy measure [10]: $E = -\sum_{i \in L} p_i \log_2 p_i$, in which p_i is the probability of the location i for the user under analysis. For each user, we also consider three locations that characterize a user's mobility: the most visited location, the second most visited location and the least visited location. Typically, the most visited location corresponds to user's home, while the second most visited location is users' work place. For each one of these locations, we evaluate the frequency of visits during the period of observation w_i, where i represents the specific location under analysis. We also define $\overline{w_i}$ as the daily average of the frequency of visits at the location i for the user under analysis. Then, we denote by w_i^{pop} the frequency of visits divided by the popularity of the location, i.e. the total frequency of the location in the dataset. In this way, we normalize the frequency of the user for a particular location considering the behaviour of all the users in the dataset. For these three locations, we also consider U_i, i.e., the number of distinct users that visited the location i in the period of observation. Out of U_i, we also compute U_i^{ratio}, in which the number of distinct users that visited the location i is divided by the total number of users in the dataset. The last measure we consider for each of the three locations is the entropy. In this case, we compute a *location entropy* E_i, that represents the predictability of a visit at the location i defined as: $E = -\sum_{u \in U_i} p_u \log_2 p_u$, where U_i is the set of users that visited the location i and p_u is the probability that a user u visited the location i. When working with trajectories, we have also a temporal information: each trajectory is composed by $\langle l_i, t_i \rangle$, in which t_i is the timestamp corresponding to time of arrival of a user at a location l_i. We exploit this information to compute the *path time* [18], i.e., the time occurring between the first and last visit of a trajectory.

Table 1. Mobility features of the individual mobility profile.

Notation	Description	Notation	Description
V	visits	$\overline{V}$	daily visits
D_{max}	max distance	D_{sum}	sum distances
D_{max}^{tot}	max distance over total max distance for a user	$\overline{D}_{sum}$	D_{sum} per day
D_{max}^{trip}	D_{max} over area	$Locs$	distinct locations
$Locs_{ratio}$	$Locs$ over area	R_g	radius of gyration
E	mobility entropy	E_i	location entropy
U_i	individuals per location	U_i^{ratio}	U_i over individuals
w_i	location frequency	w_i^{pop}	w_i over the total frequency of location i
$\overline{w}_i$	daily location frequency	PT_j	Path time per user

Privacy risk computation. The goal of this module is to compute for each user trajectory in $\mathcal{D}$ a privacy risk value by using a re-identification algorithm. We propose to apply the PRUDEnce framework (Sect. 3.1) that enables the definition and simulation of any desired privacy attacks over the entire dataset. Several attacks might be defined on the basis of the type of background knowledge possessed by an adversary [20,21]. In this paper we instantiate our risk computation using the location sequence attack, introduced in [14,15], where the adversary knows a subset of the locations visited by the individual and the temporal ordering of the visits. Given an individual u, we denote by $L(T_u)$ the sequence of locations $l_i \in T_u$ visited by u. The background knowledge category of a location sequence attack is defined as follows:

Definition 4. *Let h be the number of locations l_i of an individual u known by the adversary. The Location Sequence background knowledge is a set of configurations based on h locations, defined as $B_h = L(T_u)^{[h]}$, where $L(T_u)^{[h]}$ denotes the set of all the possible h-subsequences of the elements in the set $L(T_u)$.*

We indicate with $a \preceq b$ that a is a subsequence of b. Each instance $b \in B_h$ is a location subsequence $X_u \preceq L(T_u)$ of length h. Given a record $T \in \mathcal{D}$ we define the matching function as: $matching(T, b) = true$ if $b \preceq L(T)$, *false* otherwise. PRUDEnce uses this function to compute the probability of re-identification for any instance of background knowledge (Definition 2) enabling the privacy risk computation for each trajectory (Definition 3).

4.2 Risk Explanation Module

The last module of EXPERT is the *explainer* aiming at providing the end-user with an explanation for the predicted risk label. The objective is to increase users' awareness about the privacy risks. EXPERT is modular with respect to the explainer allowing the use of any explanation method suitable to tabular data. Since the goal is to explain a specific decision, *local* methods [11,13,22] are more suitable for this task. The main difference between them is the type of explanation returned. Lime [22] and Shap [13] are mainly based on the notion of feature importance and Lore [11] instead provides a logical rule-based explanation for the prediction. In our experiments we considered Lore and Shap as explainers. Given our ML model and an individual trajectory belonging to a user u, transformed into the mobility profile M_u and labeled with a specific privacy risk level r_u by our model, Lore (LOcal Rule-based Explanation) builds a simple, interpretable predictor by first generating a balanced set of neighbor instances of the given M_u through an ad-hoc genetic algorithm, and then extracting from such a set a decision tree classifier. A *local explanation* is then extracted from the obtained decision tree. The local explanation is a pair composed by *(i)* a *logic rule*, corresponding to the path in the tree that explains why M_u has been labeled as r_u by the predictor, and *(ii)* a set of *counterfactual rules*, explaining which changes in M_u would invert the risk class assigned. Shap (SHapley Additive exPlanations) is a local approach for interpreting model predictions that assigns to each feature an importance value for a particular prediction. Moreover, for each model's prediction Shap defines an *explanation* model. The main idea is that the explanation model is an interpretable approximation of the original model and works with simplified input data. Shap exploits the collaborative game theory to determine the importance value of a feature for the instance prediction.

5 Experiments

We experimentally validate the different components of our framework by analyzing the performance of: *i)* the prediction module implemented with different machine learning models by varying their complexity; and *ii)* the explanation module by comparing two state-of-the-art approaches.

Data. We use data containing GPS tracks of private vehicles in Tuscany (Italy) provided by Octo Telematics. We selected trajectories from an area comprising two major urban centers, Prato and Pistoia, considering the period from 1st May to 31st May 2011, for a total of 8651 distinct vehicles. We performed two different transformations of the original data in order to obtain two different datasets. In the first dataset, called `istat`, trajectory points are generalized according to the geographical tessellation provided by the Italian National Statistics Bureau (ISTAT): each point is substituted with the centroid of the geographical cell to which it belongs. We then remove redundant points, i.e., points mapped to the same cell at the same time, obtaining 2274 different locations with an average length of 31.9 points per trajectory. With respect to the second dataset, called `voronoi`, we first apply a data-driven Voronoi tessellation of the territory [1], taking into consideration the traffic density of an area, and then we used the cells of this tessellation to increase the granularity of the original trajectories. The algorithm also performs interpolation between non adjacent points[3]. We obtained 1473 different locations with an average length of 240.2 points per trajectory. For both datasets we computed the mobility features M for extracting the users' mobility profiles and the privacy risk according to the simulation of the location sequence attack (Sect. 4.1) with four background knowledge configurations B_h using $h = 2, 3, 4, 5$, getting four different risk datasets, $\Gamma_{h=2,3,4,5}$. We discretized the risk values in intervals: $[0, 0.5]$ and $(0.5, 1]$ named *low* and *high* risk class, respectively. Then, we built our classification datasets merging each risk dataset with the feature-based mobility profiles: $\langle M, \Gamma_h \rangle$, as explained in Sect. 4.1. To better handle the imbalance in the data, we learned our predictive models using stratified sampling, undersampling and 5-fold cross-validation. Tables 3 and 2 report the class balance after under-sampling the majority class. We also performed hyper-parameter tuning by grid search in the parameter space[4].

Predicting Risk. We validate the effectiveness of the prediction module of EXPERT by comparing four different ML models: Decision Tree (DT), Logistic Regression (LR), Random Forest (RF)[5], and GCFOREST (GC)[6]. Decision Tree and Logistic Regression are two well-known, white-box models. Random Forest and GCFOREST [29] are ensemble models proven to be effective when dealing with imbalanced data. This task is characterized by strong imbalance of the two risk classes, therefore being a challenging machine learning problem, where

[3] Voronoi tessellation obtained by using: http://geoanalytics.net/V-Analytics/.

[4] Hyper-parameter settings: https://github.com/francescanaretto/prp.

[5] https://scikit-learn.org/stable/.

[6] https://github.com/kingfengji/gcForest.

Table 2. Predictive models evaluation on mobility profiles derived from `istat`.

B_h	Class Balance	Under-sampling	Metric	DT	LR	RF	GC
h=2	High=77 Low=23	High=40 Low=60	$F1_{high}$	0.92 (0.00)	0.92 (0.00)	**0.94** (0.00)	**0.94** (0.02)
			P_{high}	0.90 (0.01)	0.91 (0.01)	0.91 (0.00)	**0.92** (0.01)
			R_{high}	0.93 (0.01)	**0.96** (0.00)	0.95 (0.00)	**0.96** (0.00)
			$F1_{low}$	0.69 (0.02)	0.71 (0.01)	**0.75** (0.01)	**0.75** (0.01)
			P_{low}	0.73 (0.02)	0.77 (0.01)	0.81 (0.01)	**0.82** (0.01)
			R_{low}	0.66 (0.02)	0.42 (0.03)	**0.70** (0.09)	**0.70** (0.02)
h=3	High=93 Low=7	No under-sampling	$F1_{high}$	0.96 (0.00)	0.92 (0.00)	**0.97** (0.00)	**0.97** (0.03)
			P_{high}	0.95 (0.01)	0.94 (0.01)	**0.96** (0.00)	**0.96** (0.00)
			R_{high}	0.96 (0.00)	**0.98** (0.00)	**0.98** (0.00)	**0.98** (0.00)
			$F1_{low}$	0.70 (0.02)	0.71 (0.01)	0.75 (0.01)	**0.79** (0.03)
			P_{low}	0.72 (0.02)	0.77 (0.03)	0.83 (0.03)	**0.84** (0.03)
			R_{low}	0.70 (0.06)	0.41 (0.03)	0.70 (0.04)	**0.74** (0.05)
h=4	High=95 Low=5	No under-sampling	$F1_{high}$	0.96 (0.00)	0.96 (0.00)	**0.97** (0.00)	**0.97** (0.00)
			P_{high}	0.96 (0.05)	0.95 (0.00)	0.96 (0.00)	**0.97** (0.00)
			R_{high}	0.97 (0.00)	**0.98** (0.00)	**0.98** (0.00)	**0.98** (0.00)
			$F1_{low}$	0.73 (0.02)	0.70 (0.02)	0.77 (0.02)	**0.80** (0.02)
			P_{low}	0.75 (0.02)	0.80 (0.01)	**0.85** (0.02)	**0.85** (0.09)
			R_{low}	0.70 (0.01)	0.45 (0.03)	0.74 (0.05)	**0.76** (0.03)
h=5	High=96 Low=4	No under-sampling	$F1_{high}$	0.96 (0.04)	0.96 (0.00)	**0.97** (0.00)	**0.97** (0.00)
			P_{high}	0.96 (0.04)	0.95 (0.00)	**0.97** (0.00)	**0.97** (0.00)
			R_{high}	0.96 (0.01)	**0.98** (0.00)	**0.98** (0.00)	**0.98** (0.00)
			$F1_{low}$	0.73 (0.03)	0.70 (0.03)	0.78 (0.02)	**0.80** (0.02)
			P_{low}	0.72 (0.03)	0.80 (0.05)	0.83 (0.02)	**0.85** (0.02)
			R_{low}	0.70 (0.03)	0.46 (0.03)	0.75 (0.04)	**0.76** (0.03)

the classifier performance in terms of accuracy is less significant due to the dominance of the majority class on the metric.

Indeed, as discussed in Sect. 4.1, our desiderata is a classifier with a conservative approach with respect to high risk users, to avoid their misclassification as low risk users. On the other hand, we aim at achieving high precision and recall for both high and low risk users. As a consequence, for the performance evaluation of the machine learning models, we select the following indicators: *i)* precision (P_{high}) and recall (R_{high}) on high risk; *ii)* precision (P_{low}) and recall (R_{low}) on low risk; and *iii)* the two corresponding *F1-Score* for low ($F_{1_{low}}$) and high ($F_{1_{high}}$) risk. In a setting where the size of high risk class is larger than that of the low risk one, achieving good performance for the low risk users is difficult. The results for the two datasets are shown in Tables 2 and 3. We note that `istat` represents a typical situation in the privacy context, where a high number of risky users exists. We also built `voronoi` to present a balanced situation and to verify how our models behave in such a case. In general, we found that the ensemble methods have good performance in terms of both *F1-Score* on high risk and *F1-Score* on low risk. This means that these models are suitable for our target. More precisely, we observe that, although GC and RF have comparable performance, for `istat`, that is extra imbalanced, GC performs slightly better than RF on the low risk class. Moreover, ensemble methods also outperform the white-box classifiers and again, their advantage is more evident in `istat`; especially, they considerably improve the classification scores for the more difficult category of low-risk users. Indeed, we found that GC increases of 0.04–0.06 (0.09–0.13) points the R_{low} (P_{low}) of DT and of 0.28–0.33 (0.05–0.07) points the R_{low} (P_{low}) of LR. Clearly, these results contribute to have GC with the best $F1_{low}$ for every value of h, while still maintaining a conservative behaviour highlighted by the high values of recall on high risk class (R_{high}). Regarding `voronoi`, we further notice that, although the data are more balanced, the ensemble methods always maintain the conservative approach for high risk users (high R_{high}) while improving the overall classification for low risk users ($F1_{low}$). Overall, these results suggest that GC is the most suitable option for our specific predictive task with RF as a close second one.

Table 3. Predictive models evaluation on mobility profiles derived from `voronoi`.

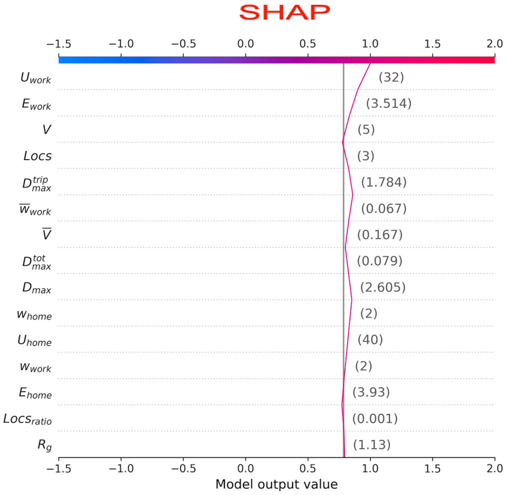
B_h	Class Balance	Under-sampling	Metric	DT	LR	RF	GC
h=2	High=28 Low=72	High=30 Low=70	$F1_{high}$	0.71 (0.02)	0.65 (0.07)	0.75 (0.02)	**0.80** (0.01)
			P_{high}	0.73 (0.01)	0.73 (0.02)	0.78 (0.01)	**0.79** (0.01)
			R_{high}	0.74 (0.04)	0.77 (0.03)	0.72 (0.02)	**0.80** (0.03)
			$F1_{low}$	0.87 (0.00)	0.86 (0.01)	**0.89** (0.01)	**0.89** (0.00)
			P_{low}	0.70 (0.01)	0.89 (0.01)	0.87 (0.01)	**0.90** (0.02)
			R_{low}	0.85 (0.01)	0.82 (0.02)	**0.91** (0.01)	0.86 (0.01)
h=3	High=55 Low=45	No under-sampling	$F1_{high}$	0.88 (0.01)	0.88 (0.01)	**0.92** (0.01)	**0.92** (0.01)
			P_{high}	0.89 (0.01)	0.88 (0.01)	**0.91** (0.00)	**0.91** (0.00)
			R_{high}	0.86 (0.02)	0.89 (0.03)	**0.92** (0.01)	**0.92** (0.01)
			$F1_{low}$	0.84 (0.02)	0.82 (0.01)	**0.87** (0.01)	**0.87** (0.01)
			P_{low}	0.80 (0.02)	0.83 (0.03)	**0.88** (0.09)	**0.88** (0.01)
			R_{low}	**0.89** (0.02)	0.81 (0.02)	0.87 (0.01)	0.86 (0.01)
h=4	High=57 Low=43	High=40 Low=60	$F1_{high}$	0.91 (0.00)	0.90 (0.00)	**0.93** (0.00)	**0.93** (0.00)
			P_{high}	0.91 (0.01)	0.88 (0.00)	0.92 (0.00)	**0.94** (0.01)
			R_{high}	0.91 (0.02)	**0.92** (0.01)	**0.92** (0.01)	0.91 (0.01)
			$F1_{low}$	0.84 (0.01)	0.80 (0.01)	**0.87** (0.01)	**0.87** (0.01)
			P_{low}	0.84 (0.03)	0.84 (0.01)	**0.85** (0.01)	**0.85** (0.01)
			R_{low}	0.84 (0.02)	0.77 (0.03)	**0.88** (0.01)	**0.88** (0.02)
h=5	High=62 Low=38	High=50 Low=50	$F1_{high}$	0.93 (0.01)	0.93 (0.01)	**0.94** (0.00)	**0.94** (0.01)
			P_{high}	0.92 (0.03)	0.90 (0.01)	0.94 (0.01)	**0.95** (0.02)
			R_{high}	0.93 (0.02)	0.93 (0.02)	**0.94** (0.01)	**0.94** (0.01)
			$F1_{low}$	0.83 (0.01)	0.80 (0.03)	**0.86** (0.01)	**0.86** (0.02)
			P_{low}	0.83 (0.03)	0.83 (0.03)	**0.86** (0.03)	**0.86** (0.02)
			R_{low}	0.84 (0.03)	0.84 (0.03)	**0.87** (0.02)	0.86 (0.03)

Explaining Risk. Regarding the explanation task in our experiments, we employed LORE [11] and SHAP [13]. We followed the experimental methodology proposed in [11]: we selected the best models from the k-fold validation presented in Sect. 5 and its associated train and test datasets. In particular, we used a RF and a GC model for $h = 2$ on the `istat` dataset. For SHAP we trained the *Kernel Explainer* on the training dataset. For LORE, we chose a genetic generation of the neighborhood and the Euclidean distance as distance among the neighbors. We performed a comparative analysis to evaluate the compactness and comprehensibility of returned explanations. To this end, we considered the diversity of the explanation structure provided by the two methods: LORE outputs rules with premises of variable lengths, while SHAP, outputs the importance of each feature in the data. Thus, we considered two different settings: i) *no-zero features*, where in the SHAP result we only keep features with importance values different from zero; and, ii) *top-k features*, that tries to automatically identify the k features with highest importance values. The value k depends on the record explanation under analysis. To detect the best k for each explanation, we used an elbow-like approach which, given the SHAP result, first sorts in descending order the importance values and then, calculates the segment s bounded by the biggest and the smallest importance values. At this point, it selects the importance value m with the maximum distance from the segment s. Thus, only features with importance values greater than or equal to m are kept. For analyzing the compactness of the explanations we considered their average lengths: LORE explanations have an average length of 2.9 ± 1.3 (RF) and 3.8 ± 1.4 (GC), against the average lenghts of paths of the decision tree of 7.8 ± 1.5. SHAP explanations have an average length of 17.1 ± 3.1(RF) and 16.2 ± 3.2 (GC) for the *no-zero features* setting, which decrease to 9.8 ± 6.3 (RF) and 8.3 ± 7.1 (GC) for the *top-k features* setting. Hence, LORE provides more compact explanations with respect to the paths of the decision tree and the SHAP importance values. We also compare the two explanation types in terms of semantic coherence. To this end, we propose to use the *Jaccard similarity* to highlight the degree of common features used for the explanations and *coherence* measure aiming at capturing the percentage of features used in LORE explanations which are important also in SHAP explanations. The *Jaccard similarity* measure, is defined as $\frac{1}{n}\sum_{i=1}^{n}\frac{F_i^{lore}\cap F_i^{shap}}{F_i^{lore}\cup F_i^{shap}}$ while the *coherence* is defined as $\frac{1}{n}\sum_{i=1}^{n}\frac{F_i^{lore}\cap F_i^{shap}}{|F_i^{lore}|}$. Here, F_i refers to the set of features included in the explanation for the record i.

Table 4 reports the results of the coherence analysis. Regarding the *no-zero features* setting, we found out that the Jaccard similarity is close to zero, highlighting that the intersection of the two feature sets is quite small compared to their union. Concerning the coherence, a value equal to 1 means that all the features of LORE are also in SHAP explanations. Results highlight that SHAP explanations contain the majority of the features used by LORE. In the *top-k features* setting, we observe a general decrease in the values of both measures. This means that the majority of the features that LORE uses in its rules are actually among the least important features of SHAP. Thus, when considering only the *top-k* features the discrepancy between SHAP important values and LORE increases. Our analysis highlights that the two methods consider different important features for providing explanations. LORE explanations tend to be more compact and easy to understand due to the logic structure of the rules. SHAP outputs a visualization and a large amount of information, which might potentially be difficult for a user to navigate. Indeed, a large number of the

Table 4. Shap vs Lore in the `istat` dataset with $h = 2$.

Setting		Jaccard	Coherence
Top-k Features	RF	0.133 ± 0.063	0.472 ± 0.381
	GC	0.096 ± 0.101	0.393 ± 0.038
No-zero Features	RF	0.133 ± 0.063	0.816 ± 0.250
	GC	0.165 ± 0.072	0.767 ± 0.232



LORE $\overline{w}^{pop}_{home} \leq 0.36, U_{home} \leq 1722, E \leq 1.09, \overline{w}_{work} \leq 0.82 \implies HighRisk$

LORE $\overline{w}^{pop}_{home} \leq 0.36, U_{home} \leq 1722, E \leq 1.09, \overline{w}_{work} \leq 0.82 \implies HighRisk$

Fig. 2. Shap vs Lore: Table 4 quantifies the similarity between the two explanations. Shap visualization (right) and the Lore rule (left) represent the explanations for a specific record classified as *high risk* by gcForest.

values of the importance features are close to zero. Moreover, given a feature used in an explanation, Lore provides a richer information that could help in understanding more about certain mobility habits that contribute to a specific risk value. For example, let us analyze Fig. 2, where we provide Shap (right) and Lore (left) explanations for a high risky user according to gcForest. With Shap a user can only understand which feature (with its specific value indicated between parentheses) is important or not for classification, while the Lore rule provides a user with a more detailed motivation, which includes the set of conditions on features that a user satisfies. For example, for the Lore explanation a user can understand that their risk depends on the fact that she travelled more than 0.09 *km* (D_{max}), their home location is visited by less than 1772 distinct users, and their work location is not enough popular in the data. This reasoning is not supported by the Shap result. After the local explanation evaluation, we also performed a comparative analysis of global feature importance among all the ML models (Table 5). An interesting result is that the number of locations (*Locs*) is the most important feature for LR, DT and GC, while for RF it is in the second position. Moreover, LR is the only one which considers the entropy of locations (home and work) as important features.

Table 5. Global top-5 most important features of machine learning models.

DT	LR	RF	GC
$Locs$ (0.45)	$Locs$ (0.35)	D_{sum} (0.15)	$Locs$ (0.07)
D_{max} (0.10)	E_{home} (0.14)	$Locs$ (0.13)	U_{work} (0.04)
U_{work} (0.06)	E_{work} (0.12)	$Locs_{ratio}$ (0.08)	$Locs_{ratio}$ (0.03)
$\overline{D}_{sum}$ (0.06)	W_{work} (0.10)	$\overline{D}_{sum}$ (0.07)	U_{home} (0.03)
U_{home} (0.06)	$\overline{D}_{sum}$ (0.08)	U_{work} (0.07)	D_{max}^{trip} (0.02)

6 Conclusions

We have presented EXPERT, a framework for predicting and explaining users' privacy risk associated to the analysis of mobility data. EXPERT exploits ML techniques that are suitable to handle extra-imbalanced data and local explainers to provide users with meaningful explanations about the predicted privacy risk. The empirical evaluation of EXPERT using real-world data demonstrate its effectiveness in predicting privacy risk and in increasing users' self-awareness in relation to potentially risky mobility behavior. The main limitation of the framework is that it requires domain expertise for extracting users' profiles for the prediction. Our future research agenda includes the substantiation of the prediction module by a ML model that does not require the extraction of mobility features. This work could also be extended to generic sequential data.

Acknowledgments. This work has been funded by the European projects SoBigData-PlusPlus (Grant Id 871042), XAI (Grant Id 834756) and HumanE-AI-Net (Grant Id 952026).

References

1. Andrienko, N.V., Andrienko, G.L.: Spatial generalization and aggregation of massive movement data. IEEE Trans. Vis. Comput. Graph. **17**(2), 205–219 (2011)
2. Armando, A., et al.: Risk-based privacy-aware information disclosure. Int. J. Secur. Softw. Eng. **6**(2), 70–89 (2015)
3. Baron, B., Musolesi, M.: Interpretable machine learning for privacy-preserving pervasive systems. IEEE Pervasive Comput. **19**(1), 73–82 (2020)
4. Cormode, G., Procopiuc, C.M., Srivastava, D., Tran, T.T.L.: Differentially private summaries for sparse data. In: ICDT 2012, pp. 299–311 (2012)
5. Craven, M., Shavlik, J.W.: Extracting tree-structured representations of trained networks. In: NIPS, pp. 24–30 (1996)
6. Craven, M.W., Shavlik, J.W.: Using sampling and queries to extract rules from trained neural networks. In: JMLR, pp. 37–45. Elsevier (1994)
7. Deng, H.: Interpreting tree ensembles with intrees. Int. J. Data Sci. Anal. **7**(4), 277–287 (2019). https://doi.org/10.1007/s41060-018-0144-8
8. Deng, M., et al.: A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements. Requir. Eng. **16**(1), 3–32 (2011). https://doi.org/10.1007/s00766-010-0115-7

9. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14
10. Eagle, N., Pentland, A.S.: Eigenbehaviors: identifying structure in routine. Behav. Ecol. Sociobiol. **63**, 1057–1066 (2009). https://doi.org/10.1007/s00265-009-0739-0
11. Guidotti, R., et al.: Factual and counterfactual explanations for black box decision making. IEEE Intell. Syst. **34**(6), 14–23 (2019)
12. Guidotti, R., et al.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**, 1–42 (2019)
13. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: NIPS, pp. 4765–4774 (2017)
14. Mohammed, N., et al.: Walking in the crowd: anonymizing trajectory data for pattern analysis. In: CIKM, pp. 1441–1444. ACM (2009)
15. Monreale, A., et al.: Movement data anonymity through generalization. TDP **3**(2), 91–121 (2010)
16. Monreale, A., et al.: Privacy-preserving distributed movement data aggregation. In: Vandenbroucke, D., Bucher, B., Crompvoets, J. (eds.) Geographic Information Science at the Heart of Europe. Lecture Notes in Geoinformation and Cartography. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-00615-4_13
17. de Montjoye, Y.A., et al.: Unique in the crowd: the privacy bounds of human mobility. Sci. Rep. **3**, 1376 (2013)
18. Muntean, C.I., et al.: On learning prediction models for tourists paths. ACM Trans. Intell. Syst. Technol. **7**(1), 8:1–8:34 (2015)
19. Pappalardo, L., et al.: Returners and explorers dichotomy in human mobility. Nat. Commun. **6**, 1–8 (2015)
20. Pellungrini, R., et al.: A data mining approach to assess privacy risk in human mobility data. ACM TIST **9**(3), 31:1–31:27 (2018)
21. Pratesi, F., et al.: Prudence: a system for assessing privacy risk vs utility in data sharing ecosystems. Trans. Data Priv. **11**(2), 139–167 (2018)
22. Ribeiro, M.T., et al.: "Why should I trust you?": explaining the predictions of any classifier. In: ACM SIGKDD, pp. 1135–1144 (2016)
23. Rossi, L., Musolesi, M.: It's the way you check-in: identifying users in location-based social networks. In: COSN, pp. 215–226. ACM (2014)
24. Samarati, P., Sweeney, L.: Generalizing data to provide anonymity when disclosing information (abstract). In: PODS, p. 188. ACM (1998)
25. Song, Y., et al.: Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. In: International Workshop on Privacy-Preserving IR: When Information Retrieval Meets Privacy and Security, pp. 19–24 (2014)
26. Terrovitis, M., Mamoulis, N.: Privacy preservation in the publication of trajectories. In: MDM, pp. 65–72 (2008)
27. Zhang, Y.L., et al.: Distributed deep forest and its application to automatic detection of cash-out fraud. ACM Trans. Intell. Syst. Technol. **10**(5), 1–9 (2019)
28. Zheng, Y.: Trajectory data mining: an overview. ACM TIST **6**(3), 29:1–29:41 (2015)
29. Zhou, Z.H., Feng, J.: Deep forest: towards an alternative to deep neural networks. In: IJCAI, pp. 3553–3559 (2017)