Driver risk identification using telematics and contextual data crossed with previous claims history

Sarah Ben Slimene, Mounîm A. El-Yacoubi

Driver Risk Identification using Telematics and contextual data crossed with previous claims history

Sarah Ben Slimene
Telecom SudParis
Institut Polytechnique de Paris
Palaiseau, France
sarah.benslimene96@gmail.com

Mounim A. El-Yacoubi
Telecom SudParis
Institut Polytechnique de Paris
Palaiseau, France
mounim.el_yacoubi@telecom-sudparis.eu

***Abstract*—We propose, in this paper, a comprehensive study on risk assessment related to car insurance, based on claim and telematics data, collected from a dataset of voluntary drivers. Our work addresses experimental settings not covered before in the state of the art, such as the collection of telematic data within a period significantly after the claim reporting one, and coping with much more unbalanced data owing to the rarity of the claim class. To address these issues, we propose weighted XGBoost models that model the telematic-based features, represented as tabular feature heatmaps and mitigate the class unbalanced problem. Our heatmaps encore not only speed-acceleration distributions, but also speed-jerk (acceleration derivative) distributions, not considered in previous studies. To ensure model interpretability, we assess the importance of each feature in order to reveal the telematic features that are most discriminant for claim detection. Owing to rarity of the claim data, we propose also one-class machine learning models, namely 1-class SVM and Isolation Forests, trained only on the most dominant class, the samples from the underrepresented one being assessed only in the test phase and considered as outliers or anomalies. In addition, we propose a novel unsupervised machine learning strategy consisting of a two-stage clustering scheme that not only allows to infer a driver signature and estimate, through sound information-theoretic measures, how stable a driver behavior is, but also uncovers the correlation between driver groups (clusters) and claim distributions. The first clustering stage uncovers behavioral clusters, i.e. monthly-based heatmap prototypes, irrespective of drivers, while the second uncovers yearly based driver clusters with similar behavior w.r.t the first-level clusters. We obtain promising results given the adverse conditions considered and we provide guidelines in the conclusion for developing more effective driving risk assessment based on telematic data.**

***Index Terms*—Car insurance, Claim detection, Telematics, Machine Learning, Interpretability**

I. Introduction

A. Context and motivation

Driving insurance is a fundamental component of modern society as it provides protection and peace of mind to individuals and businesses alike. Traditional methods for assessing risk and setting premiums have relied, for many years, on policies such as bonus-malus. Early studies on risk assessment in the insurance industry have relied on traditional risk factors such as age, gender, and driving record to determine premium pricing. These risk factors, however, are too broad and do not accurately reflect individual driving behavior, leading to overpricing for some individuals and underpricing for others. According to a recent literature review [1] on risk assessment in insurance industry, such risk factors are limited in their ability to accurately predict individual risk.

To enhance risk assessment and pricing strategies, the insurance industry has been recently exploring cutting-edge technologies to collect telematics data as they provide a comprehensive view of driver behavior. By leveraging smart devices like GPS trackers, smartphones, and in-vehicle embedded sensors, telematics can capture key information such as speed, acceleration, braking, and location. Telematics can also provide additional contextual data such as driven distance, day time, and weather conditions, which can further improve risk assessment. This technology has given rise to alternative insurance models like Pay-As-You-Drive (PAYD) or usage-based insurance (UBI) [27], whereby policyholders install telematics devices in their vehicles and receive premium discounts in return. Telematics technology, therefore, enables insurance companies to personalize pricing and risk assessment by adjusting premiums based on policyholders' driving behavior and usage, ensuring thereby a personalized pricing model based on individual risk profiles. This new policy making has been boosted recently by French regulations which, in September 2021, gave the green light to telematics data in car insurance pricing, subject to guidelines and regulations set and controlled by the French data protection authority [9]. In the telematics-based insurance literature, safety variables related to road accidents have been identified as key factors for assessing personalized risk and providing a more comprehensive coverage of driving behavior, leading to a reduction in the frequency and severity of claims [18] [17] [30].

Numerous studies have investigated the relationship between mileage and accident risk. Verbelen et al. (2018) [37], for instance, considered multivariate exposures as compositional variables, while Paefgen et al. (2014)[31] grouped mileage based on various factors. Ayuso et al. (2016b) [4] [2] used a Weibull regression in a survival model to examine the distance traveled before a first accident caused by the driver, revealing that night driving and speeding are risk factors of an accident. Gender differences in accident risk were explored by Ayuso et al. (2016a), who found that they were primarily due to variations in vehicle usage, with men driving more frequently than women. Ayuso et al. (2019) [3] further

incorporated data on driving habits, such as the percentage of distance driven at night, above speed limit, or in urban areas. Lemaire et al. (2016) [25] found mileage to be the most significant predictor of the number of at-fault claims based on data from a Taiwan insurer. Finally, Boucher et al. (2017) [6] examined the non-linear effects of duration and distance exposure on accident risk using a generalized additive model.

Recent studies on behavioral telematics can be categorized in two ways, according either to the type of tasks they are targeting, or to the type of risk assessment methods they rely on. In task-based categorization, we distinguish mainly three categories: Classification to predict claims occurrence [33], Estimation of claims frequency [38], [12](2022), [11] [15], [41], [14], and Classification to predict claims based on past claims frequency. The Second categorisation is based on the type of models used for risk assessment, namely actuarial models, Machine Learning (ML) models, and Deep Learning (DL) models. One popular method used in actuarial science is Generalized Linear Models (GLMs), used to model the frequency of claims using risk factors and telematics data. Some use telematics data to create speed-acceleration heatmaps (probability distributions), first introduced in [38] to identify driving styles and predict claims frequency. In addition to speed-acceleration heatmaps, these studies also utilized the encoded data to extract features and model claims frequency using GLMs and covariates. These models were used to identify driving behavior patterns and to predict the likelihood of claims. Similar studies, including [14] [11] [15], [39], have demonstrated that analyzing telematics data in this manner can effectively recognize driver behavior patterns and anticipate potential claims. Regarding machine learning (ML) approaches for auto insurance modeling, various methods, such as logistic regression, XGBoost, random forest, decision trees, naive Bayes, and KNN, have been proposed to predict claim occurrence [21]. A Tree-based ML method has been proposed in [22] to uncover insights in insurance tariff plans. The ML category has gained increasing attention recently as illustrated by the review for pricing and reserving in [5] that reviews the applications of ML to actuarial science. The deep learning category is not yet as popular since it usually requires modeling directly the raw data which is not straightforward given the huge spatio-temporal data streams associated with telematic data. To overcome this issue, a DL-based model has been proposed in [13] by proposing a Convolutional Neural Network (CNN) that takes as input images obtained by aggregating the telematic data into heatmaps. The studies above have shown that GLMs can provide promising results in modeling claims frequency. These models, however, require human actuarial expertise to identify the relevant variables and the linearity assumptions between variables may not always hold. ML models, by contrast, do not require expertise knowledge in actuarial science and they are able to learn complex non-linear mapping functions between the input and the output, although they rely on features extracted in a handcrafted way from telematic data. Despite the potential of Deep learning models to overcome this issue by modeling the raw data, the huge spatio-temporal resolution of the long telematic input stream and the large noise it exhibits make such a modeling a challenging task.

B. Proposed work

Our proposed work lies in the category of past claims-based prediction by leveraging telematic data and machine learning. The state of the art is very limited for this category as very few studies seek to correlate driving behavior to claims not occurring necessarily in the driving period but prior to this period. Thus, such telematic data-past claims association is hard to find as past driver experience may not be significantly correlated with future claims. An accident, for instance, may change significantly driver behavior to avoid subsequent risks. From this viewpoint, our work is similar to [15], which also models past claims. However, our work is different in several ways. First, in [15], the driving period considered in the experiments was three months from 01/05/2016 to 31/07/2016, with reported claims from 01/01/2014 to 29/05/2017, while in our experiments, the driving period was from December 2020 to December 2022, with reported claims from 2016 to 2021, with only 20% of the claims occurring during the two year driving period and 60% occurring more than two years earlier. This means that, on average, a claim in our study occurs a large period before the driving experiment starts, which may decrease the correlation between telematics and claims. Also, in [15], the average claims frequency is 0.24 per year per car/driver, while in our data, this ratio is 0.011. As pointed out in [15], the ratio of 0.24 is much higher than typically in Europe. This makes our data much more unbalanced in terms of claims occurrence w.r.t non-claims. An additional difference is that, while [15] propose to estimate claims' frequency based on actuarial models, our task is to classify a telematics-based driver behavior into two classes: (past) claim or none. While these facts show that our data are extremely adverse to any machine learning model, this reflects the type of data that can be collected for such studies from random voluntary drivers. The considerations above have guided us to consider ML models that are suitable for classifying telematic data aggregated into heatmaps in a similar way as the state of the art. Our heatmaps, however, encore not only speed-acceleration distributions, but also speed-jerk distributions, which has not been considered in the previous studies. Jerk corresponds to the derivation of the acceleration and thus encodes the hesitations or acceleration-deceleration transitions that may characterize some drivers. As these aggregated telematic data generate tabular features, we propose XGBoost models that not only have been demonstrated to outperform DL models for such kind of input [16], but also have good ability to cope with unbalanced class distribution through their weighted training model version. To ensure the explainability of our models, we estimate the importance of each feature in order to reveal the telematic features that are most discriminant for claim detection. Although our XGBoost models show promising results, the very limited amount of claim data and the severe unbalanced class distribution issue it entails have guided us

to propose one-class supervised ML models, namely one-class SVM and Isolation Forests, trained only on the most dominant class, the underrepresented one being assessed only in the test phase and considered as an anomaly class. For both classification schemes, we use as evaluation metrics, AUC (Area under the Curve), $F1$ score and $BalancedAccuracy$ which is more reliable for unbalanced class distribution settings. To cope with the unbalanced class distribution problem, we propose also a novel unsupervised ML technique consisting of a two-stage clustering scheme that not only allows inferring a driver signature and estimate, through sound information-theoretic measures, how stable a driver behavior is, but also uncovers the correlation between driver groups (clusters) and claim distributions. The first clustering stage uncovers behavioral clusters, i.e. monthly-based heatmap prototypes, irrespective of drivers, while the second uncovers yearly-based driver clusters with similar behavior distribution over the first-level clusters.

To summarize, our contributions are as follows: 1) We extract a wide range of features from the raw second-by-second telematic data by considering longitudinal and transversal speed, acceleration and jerk, aggregated at the year, month and week levels; 2) We propose a two-class weighted XGBoost that accommodates the heavily unbalanced class distribution; 3) We propose a one-class supervised scheme that does not require training on the underrepresented class (claims) and treat claim detection as an anomaly detection problem; 4) We propose a novel two-stage unsupervised model that not only infers driver signature but also correlates driver groups with claim distribution through information-theoretic measures; 5) we estimate the importance of the extracted features for claim detection to make our models explainable.

Our paper is structured as follows. Section 2 describes the data and details the feature extraction process. Sections 3 and 4 present the details of, respectively, the XGBoost and one-class models, used for classification along with the experiments and the results obtained. Section 5 details our two-stage unsupervised scheme proposed to infer driver signature and to correlate driver groups with claim distribution. Section 6 concludes the paper.

II. Methodology

A. Dataset

Our telematic dataset consists of GPS location data, speed, acceleration, in all directions, collected on a second-by-second basis from smart devices installed on cars of voluntary drivers. A total of 1032 drivers/cars have participated in the study and their telematic data were collected over up to two years of driving experience from December 2020 to December 2022, generating over +10 M second-per-second data. The speed, x acceleration (longitudinal), and y acceleration (horizontal) ranges are, respectively, [0, 250] km/h, [-1400, 1400] mg and [-1000, 1100] mg. Based on the instantaneous acceleration at time t, $A(t)$, we extract the instantaneous jerk at time t, $J(t)$. The x and y jerk at time are defined as:

$$J_x(t) = A_x(t+1) - A_x(t) \quad (1)$$

$$J_y(t) = A_y(t+1) - A_y(t) \quad (2)$$

The jerk encodes the rate of acceleration change and thus encodes driver hesitations related, for instance, to traffic jam, or unexpected events. It may, therefore, be relevant for risk assessment.

B. Telematics Heatmaps-based Feature Extraction

From these raw data, we extract global features by aggregating the instantaneous values into buckets, i.e. rectangular intervals, where a horizontal interval corresponds to a specific speed range, and a vertical interval corresponds to, respectively a specific acceleration interval or jerk interval. For speed, hereafter referred to by V (velocity), based on the following segments [0:30], [30:90],[90:130] associated, in France, to Urban zone, Extra Urban zone, and Motorway zone respectively, we consider seven finer intervals for the generation of our heatmaps. The intervals are defined by these thresholds: [0, 30, 60, 90, 110, 120, 130, 250]. Likewise we generate 16 intervals for A_x, A_y, J_x and J_y.

A_x : [-1400, -850, -650, -450, -350, -250, -150, -50, 0, 50, 150, 250, 350, 450, 650, 850, 1400]

A_y: [-1400, -850, -650, -450, -350, -250, -150, -50, 0, 50, 150, 250, 350, 450, 650, 850, 1200]

J_x and J_y : [-1000, -450, -400, -350, -300, -250, -150, -50, 0, 50, 150, 250, 300, 350, 400, 450, 1100]

According to these segmentations, we generate four heatmaps : $V - A_x$, $V - A_y$, $V - J\_x$, and $V - J\_x$, by aggregating the associated buckets over one week, one month, or one year. The dimension for each map for each duration is 112 as the horizontal axis consists of 7 intervals and the vertical axis always consists of 16 intervals. The reason for this gross aggregation is threefold. First, it allows reducing the huge dimensionality associated with second-by-second telematic data, that is unmanageable directly. Second, such reduction allows the use of machine learning models which otherwise would not be possible as, in this case, the dimension would be much larger than the number of observations. Third, it allows to mitigate the huge noise inherent to embedded sensors on cars, and the large non-stationarity of the telematic data stemming from various factors such as daily route state and weather conditions, traffic jam levels, week days, etc.

III. Past claims frequency detection

We describe in this section our claim classification approach. After detailing how we encore the claim data, we describe our weighted XGBoost model and present the experiments and the results obtained.

A. Encoding of Claim Data

Table I shows the distribution of claims in the training and test datasets, at the yearly, monthly and weekly driving periods (aggregation levels). The table shows the number of driving periods and their distribution at different levels of claim responsibility. In the context of insurance, the *degree* of driver responsibility refers to the degree to which an insured party is at fault or responsible for a claim. It is attributed the value

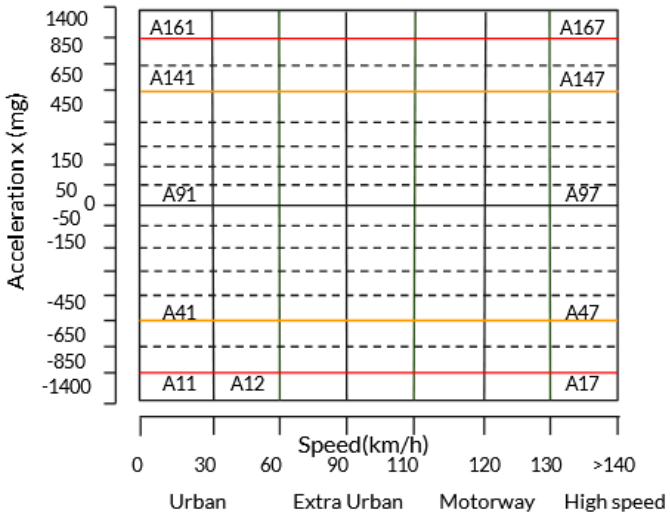

Fig. 1: Caption

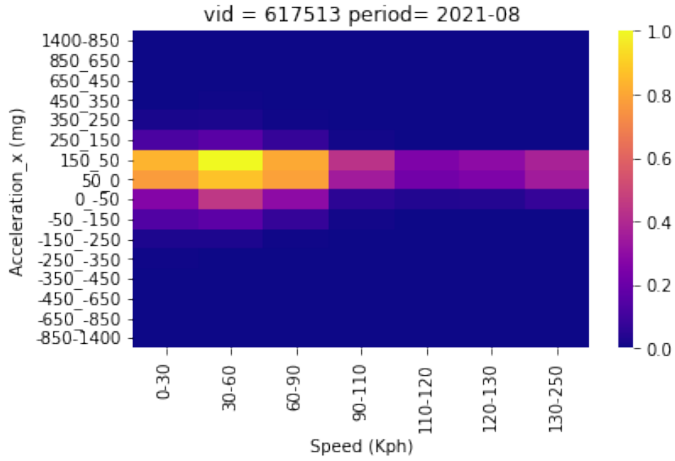

Fig. 2: Heatmap

of 1, 0.5 or 0 depending on whether the driver has total, 50% or no responsibility, respectively. The *level* of responsibility we consider in this study refers to the accumulated degree of responsibility over the six year report claim period. Please note that drivers with no claims and those with claims but involving null responsibility are merged into one class (Level of Responsibility = 0). According to this encoding scheme, we obtain in our dataset seven claim levels of responsibility $lr(i)$ for i = 0, 1, ...6, with $lr(i) = (0, 0.5, 1, 1.5, 2, 2.5, 3.5)$. Owing to the severe under-representability of the positive levels of responsibility, we merge them into one class: the claim class. In the table, the count regarding the merged claim class corresponds to the symbol $\mathbb{R}^+$ (i.e. $lr(i) > 0$). Despite this merging, the claim class remains heavily underrepresented as it account for less than 9% of the driving periods.

B. XGBoost models: $V - A_x$ / $V - A_y$ / $V - J_x$ / $V - J_y$

Although the trend in machine learning for prediction tasks on large datasets is to use deep neural networks as they are able to model complex non-linear hierarchical mappings between the input and the target output, it has been shown recently [16] that tree-based models, and XGBoost in particular, outperform deep models when the input data are in the form of tabular features. Our input telematics-based heatmaps are indeed tabular data and are characterized by significant uninformative features and irregular patterns in the input-target mapping as a result of the large time gap that might exist between the driving period and the claim period, and the non-availability of key contextual features such as the speed limit in different driving zones. Owing to these reasons, we propose XGBoost models [7] that not only have been demonstrated to outperform DL models for such kind of input, but also have good ability to cope with unbalanced class distributions through their weighted training model version. XGBoost is a scalable and fast implementation of gradient boosting, that consists of training an ensemble of shallow decision trees, where each iteration harnesses the error residuals of the previous tree model to train the next one. It can be used for classification or regression, and the final prediction is a weighted sum of all the tree predictions. Weighted XGBoost assigns to the samples weights that are inversely proportional to their class frequency in the training set. As the claim data represent less than 9% in the training data, these samples get much higher weights in the XGBoost loss function than the normal data do.

C. Model Evaluation

1) Tuning of the XGBoost hyperparameters : The training of XGBoost consists of jointly fitting the trees and optimizing the model hyperparameters. Chief among the latter are the number of trees to fit, the fraction of features to be randomly sampled for each tree, and the maximum depth of a tree, but there are other important hyperparameters as well such as the learning rate and the regularization terms. Each hyperparameter has a specific range of possible values chosen according to the XGBoost implementation, with the optimal value being dependent on the characteristics of the dataset at hand. Optimizing these hyperparameters is key to achieve optimal performance as this will ensure the best tradeoff between underfitting and overfitting. To perform such an optimization, we adopt a 3-fold cross-validation scheme, by dividing the training dataset into three equal subsets. It is worth noting that the split is performed in a stratified way to keep the same distribution in terms of claims' frequency over the three subsets, in order to avoid optimisation bias. For each hyperparameter combination, each subset is used as a validation fold to evaluate the model while the remaining two-thirds are used to fit the model trees, the overall evaluation performance being the average performance over the three validation subsets. By trying different combinations of the hyperparameters using a grid search, we can find the optimal combination that produces the best average performance on the three validation subsets. Once this combination is found,

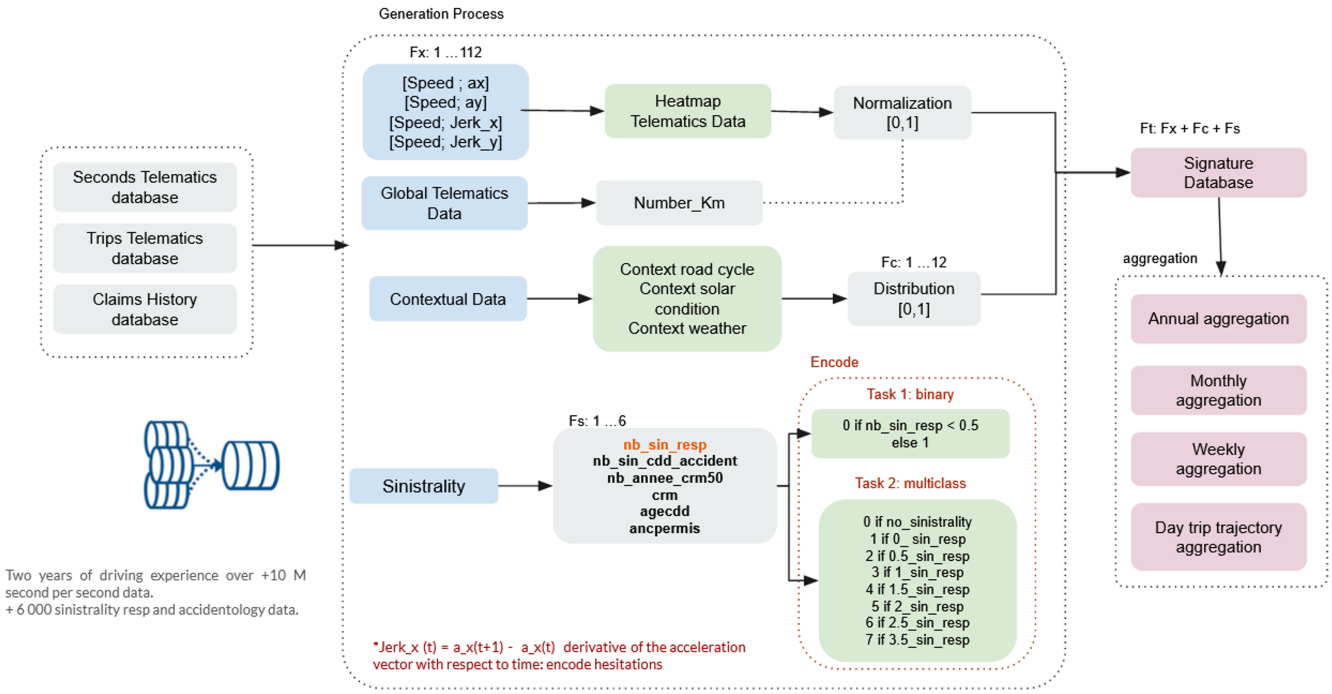

Fig. 3: Drivers signature generation process: row data preprocessing

TABLE I: Aggregation Distribution

Aggregation Level	Train	Test
Annual	$\{0:1171\}$ $\{0.5\,\|\,87, 1.0\,\|\,4, 1.5\,\|\,2, 2.0\,\|\,10, 2.5\,\|\,6, 3.5\,\|\,1, \mathbb{R}^+\,\|\,110\}$	$\{0:507\}$ $\{0.5\,\|\,34, 1.0\,\|\,2, 1.5\,\|\,2, 2.0\,\|\,4, 2.5\,\|\,1, 3.5\,\|\,2, \mathbb{R}^+\,\|\,45\}$
Monthly	$\{0:8207\}$ $\{0.5\,\|\,626, 1.0\,\|\,25, 1.5\,\|\,7, 2.0\,\|\,63, 2.5\,\|\,40, 3.5\,\|\,8, \mathbb{R}^+\,\|\,769\}$	$\{0:3656\}$ $\{0.5\,\|\,215, 1.0\,\|\,18, 1.5\,\|\,14, 2.0\,\|\,33, 2.5\,\|\,12, 3.5\,\|\,21, \mathbb{R}^+\,\|\,313\}$
Weekly	$\{0:29378\}$ $\{0.5\,\|\,2248, 2.0\,\|\,222, 2.5\,\|\,155, 1.0\,\|\,89, 1.5\,\|\,18, 3.5\,\|\,18, \mathbb{R}^+\,\|\,2750\}$	$\{0:13037\}$ $\{0.5\,\|\,759, 1.0\,\|\,71, 2.0\,\|\,102, 3.5\,\|\,85, 2.5\,\|\,52, 1.5\,\|\,41, \mathbb{R}^+\,\|\,1110\}$

we train, with the optimal hyperparameters, a new XGBoost model on the whole training dataset, as splitting the training dataset is no longer needed, and it is important to leverage the whole training data available for fitting the trees.

2) Results : Table II shows the results obtained with our weighted XGBoost algorithm based on the different Telematics-based features proposed in this study, namely, $V-A_x$, $V-A_y$, $V-J_x$, and $V-J_y$, heatmaps, extracted at the year, month and week levels. The sizes of the training|test datasets at the year, month and week levels are 1281|552, 9012|3988, and 32128|14147, respectively. As the split of the training and test datasets are done in a stratified way, all the sets above comprise the same ratio between the claim and normal data, roughly 9%. The results are expressed in terms of the $Balanced Accuracy$ (BA), Area under the Curve (AUC), $Precision$, $Recall$, and the $F1$ score. These metrics are defined as follows:

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

$$Recall = Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$BA = \frac{Sensitivity + Specificity}{2}$$

$$AUC = \sum_{i=1}^{n-1} \frac{(TPR_i + TPR_{i+1}) \times (FPR_{i+1} - FPR_i)}{2}$$

where FPR_i represents the false positive rate and TPR_i represents the true positive rate at the i-th threshold.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Specificity is the proportion of correctly predicted negative instances (Class 1: normal) out of all actual negative instances. A high Specificity indicates a model with low rate of falsely classifying normal as actual claims. $Precision$ is the proportion of correctly predicted positive instances (Class 1: claims) out of all instances predicted as positive. A high $Precision$ indicates a model with low rate of falsely predicting claims

when the instances are actually normal. $Recall$ (also known as $Sensitivity$ or true positive rate) is the proportion of correctly predicted positive instances (Class 1: claims) out of all actual positive instances. A high $Recall$ indicates a model with low rate of falsely classifying actual claims as normal. The $F1$ score is the harmonic mean of $Precision$ and $Recall$, providing a single metric that balances both measures. The $F1$ score is high when both $Precision$ and $Recall$ are high, indicating a good balance between correctly predicting claims and avoiding false positives and false negatives. BA is a reliable measure when there is large unbalance distribution between negative class (here, normal: $\tilde{9}2\%$) and the positive class (here, claims: $\tilde{9}\%$). Tu understand this, let us consider the traditional $Accuracy$ metrics, defined as the proportion of well classified instances irrespective of their class (normal or claims). Assume we consider a blind classifier that always assign a test instance to the dominant class (normal). In such a setting, the accuracy obtained is 92%, a high rate that is nonetheless misleading as the classifier never detects a claim. By contrast, BA in this case is merely 50% as Specificity = 100% but $Recall$ = 0%. Thus, BA has reliable interpretation to assess how a model behaves w.r.t a blind classifier, while taking the under-represented (rare) class on equal footing with the dominant one. In our presentation of the results, we will emphasize mainly BA even though we report also AUC and $F1$ as they are also popular metrics of the state of the art.

As observed in Table II, the best results, in terms of BA, are obtained at the year level with A_x and A_y and are about 62%. This is a very promising result given the rarity of the claim class, with a significant improvement of 12% over the blind classifier. J_x and J_y also bring improvements but less significantly. J_x is 3% better than J_y, reflecting the fact that longitudinal (driving direction) acceleration-deceleration hesitations are more correlated to claims. At the month levels, the performance with Jerk becomes better than with acceleration as the latter is behaving merely as a blind classifier. This may reflect the fact that the monthly acceleration heatmaps become so noisy and unstable that their correlation with claims becomes low. Jerk, it turns out, does keep some correlation with claims but to a significantly lower extent w.r.t the year level. The instability stems for underlying factors, such a possible vacation periods and special events, not taken into account in the study. At the week level, both acceleration and jerk provide performance similar to a blind classifier, hinting to their high noisiness and inability to capture relevant information for detecting claims. The reasons why some performances at the month and week level are poor lie in the fact that considering heatmaps at such time resolutions introduces high amount of noisiness that might stem from driver behavior change due to seasonal changes, vacation periods, or whether the car was driven mainly by a different driver. Such changes are usually dissipated over the annual period.

Although the models based on year aggregation suffer less from the factors above, they still operate under adverse effects like the rarity of the claim class, and the non-availability of key contextual information such as speed limit. As an example, a speed of 60 km/h is normal in an extra-urban zone while it far exceeds the limit in an urban zone (30 km/h). If the speed limit were available, considering, as feature, speed difference w.r.t the limit instead of absolute speed would have likely lead to significantly better results. Despite these negative factors, our model was able to provide promising results by detecting a significant number of claims. Concretely, the A_x based model has a $Recall$ of 73% as it detects 33 claims over 45. An instance is classified as a claim if the XGBoost output probability score is higher than the threshold of 0.5. We may be tempted to change this threshold and retain the one maximizing the $Recall$, under the constraint that the Specificity does not decrease significantly. We have carried out such an experiment and we report the results in Figure 4. As observed, lowering the threshold doe increase the $Recall$, allowing it even to reach 100% for a threshold ≤ 0.35, but this comes at the expense of a sharp decrease of Specificity. The figure shows, in particular, that 0.5 is the threshold allowing for the optimal trade-off between Specificity and $Recall$ as reflected by the maximum BA obtained.

A natural strategy to seek further improvements of the results above is to fuse the Acceleration and Jerk based models. This fusion can be score-based or feature-based. The former consists of fusing the scores output by the combined models while the latter consists of concatenating the feature vectors to fuse and input them to a single XGBoost model. The bottom part of Table II shows such results when combining A_x and A_y. As shown, whether the fusion is score-based or feature-based, no improvement is obtained for the different driving periods considered. The other fusion experiments, considering two or more feature types, have not led to improvements either and are not shown in the table. This may be explained mainly by the small number of claims in the test set and by the fact that a large number of them is already detected prior to fusion (Recall = 33/45 with A_x), and that the remaining ones may be due to other factors such as the absent of the context, etc. These are preliminary results and other more sophisticated fusion schemes [26] [10] [24] might be considered in the future, by training, for instance, a neural network that takes as input all the features types and learn how to combine them in a non-linear way. Prior to this, nonetheless, a large number of claims should be added to the dataset, to measure the effect of fusion and other enhancement strategies.

D. Shap Values

The SHAP (SHapley Additive exPlanations) [29] explanation technique is recent method that allows for interpreting the predictions of machine learning models, considered as black boxes, by quantifying the impact of individual features on the global model predictions, or on the model's prediction for a specific instance. These feature contributions are referred to as the SHAP values. This knowledge may empower us to make informed decisions, identify influential factors, detect potential biases, and improve the transparency and trustworthiness of our model's outputs. In this section, we show our assessment of the importance of each feature input to our XGBoost model.

TABLE II: Accuracy metrics obtained with the weighted XGBoost model. SF stands for Score Fusion while FF stands for Feature Fusion

	BA	AUC	$Precision$	$Recall$	$F1$
A_x year	0.615	0.576	0.95	0.50	0.65
			0.11	0.73	0.20
A_y year	0.619	0.563	0.97	0.39	0.56
			0.11	0.84	0.19
$Jerk_x$ year	0.592	0.591	0.95	0.45	0.61
			0.11	0.73	0.19
$Jerk_y$ year	0.560	0.571	0.94	0.43	0.59
			0.10	0.69	0.17
A_x month	0.522	0.549	0.93	0.78	0.85
			0.09	0.27	0.14
A_y month	0.481	0.475	0.92	0.80	0.86
			0.07	0.16	0.09
$Jerk_x$ month	0.547	0.571	0.93	0.61	0.74
			0.10	0.49	0.16
$Jerk_y$ month	0.547	0.573	0.93	0.76	0.84
			0.11	0.34	0.16
A_x week	0.524	0.551	0.93	0.84	0.88
			0.10	0.21	0.14
A_y week	0.484	0.476	0.92	0.73	0.82
			0.07	0.24	0.11
$Jerk_x$ week	0.549	0.563	0.94	0.44	0.60
			0.09	0.66	0.16
$Jerk_y$ week	0.515	0.522	0.92	0.69	0.79
			0.09	0.34	0.14
SF (A_x, A_y) year	0.578	0.566	0.94	0.47	0.63
			0.10	0.69	0.18
FF (A_x, A_y) year	0.518	0.564	0.92	0.59	0.72
			0.09	0.44	0.15
SF (A_x, A_y) month	0.492	0.509	0.92	0.84	0.88
			0.07	0.14	0.09
FF (A_x, A_y) month	0.521	0.532	0.93	0.66	0.77
			0.09	0.38	0.14
SF (A_x, A_y) week	0.49	0.496	0.92	0.86	0.89
			0.07	0.12	0.09
FF (A_x, A_y) week	0.503	0.542	0.92	0.90	0.91
			0.08	0.11	0.09

To understand which feature has significant importance, we compute the average future contribution as $Avg = 1/112$, as feature dimension = 112. A feature importance is expressed in terms of how much it is higher than Avg. In our dataset, the acceleration is measured in mg, g being gravity acceleration. To ease interpretabilty, we define three acceleration types: Soft, Medium and Intense. Concretely, we categorize Braking intensity and Acceleration into different levels. Soft braking is characterized by an acceleration between 0 and -450 m/s^2, Medium braking by the range of -450 to -650 m/s^2, and Intense braking by the range of -650 and -1400 m/s^2. Similarly, Soft acceleration refers to a gradual increase in speed, with an acceleration range from 0 to 450 m/s^2, Medium acceleration, with a range from 450 to 650 m/s^2, and Intense acceleration with a range from 650 to 1400 m/s^2. Each result is associated with a specific speed , acceleration/ braking combination of Heatmap features values.

Figure 5 shows the 20 most important features, from top to bottom in decreasing order, according to their *global* impact on the model output on the normal/claim prediction task on the test set, when considering the $V - A_x$ feature map consisting of 112 features overall. The vertical position of each feature

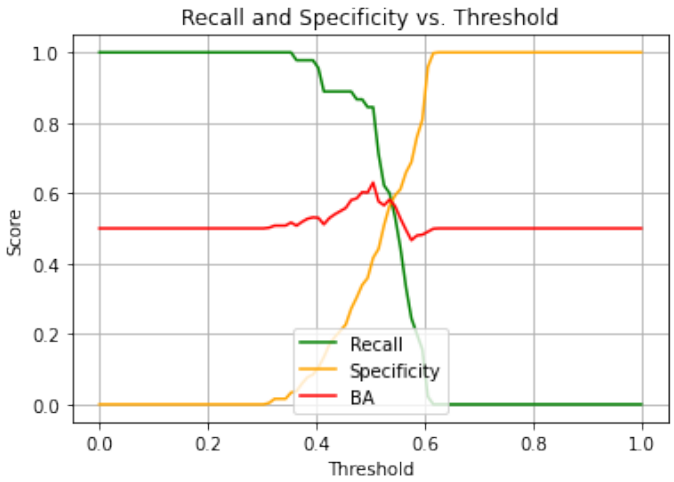

Fig. 4: $Specificity$, $Recall$, and $Balance\ Accuracy$ as a function of the decision threshold

correspond to its rank among the best features. This is reflected by the global SHAP value along side it (e.g. top feature: -A50:0-S60-90 : 0.1). Associated with each feature name is a horizontal bar, with colors represented in shades between blue and red, corresponding to the values that this feature has taken on the different instances on the test set. High positive values correspond to bright shades of red, while high negative values correspond to bright shades of blue. Darker shades of red and blue correspond to moderate positive and negative values respectively. Associated with these feature values are their SHAP values, i.e. their importance or contribution on the model prediction. These SHAP values, shown on the horizontal axis, represent the direction and magnitude of the features' impact on the model's output. A positive feature value v indicates that it has a positive contribution of v towards predicting the claim class, and vice versa. When a feature has a negative SHAP value, this means that the feature value is associated with a decrease in the model prediction of the claim class, albeit this does not necessarily imply that the normal class will be output as this depends on the other features as well. Similarly, when a feature has a positive SHAP value, this means that the feature value increases the model prediction of the claim class, albeit this does not guarantee the prediction of this class. As far as the amplitude is concerned, the larger (farther from the origin) the SHAP value, the higher the impact of the feature value on the model output.

According to Figure 5, we draw the following interpretations for the 10 top features:

- '-A50:0-S60-90' indicating soft braking with a speed range of 60-90 km/h is the most important feature overall. Its SHAP value, 10%, has more than 10 times feature importance, w.r.t Avg, for claim prediction.
- 'A350:250-S60-90' corresponds to soft acceleration with a speed range of 60-90 km/h (SHAP = 9%).
- '-A350:250-S60-90' denotes soft braking with a speed range of 60-90 km/h (SHAP = 9%).

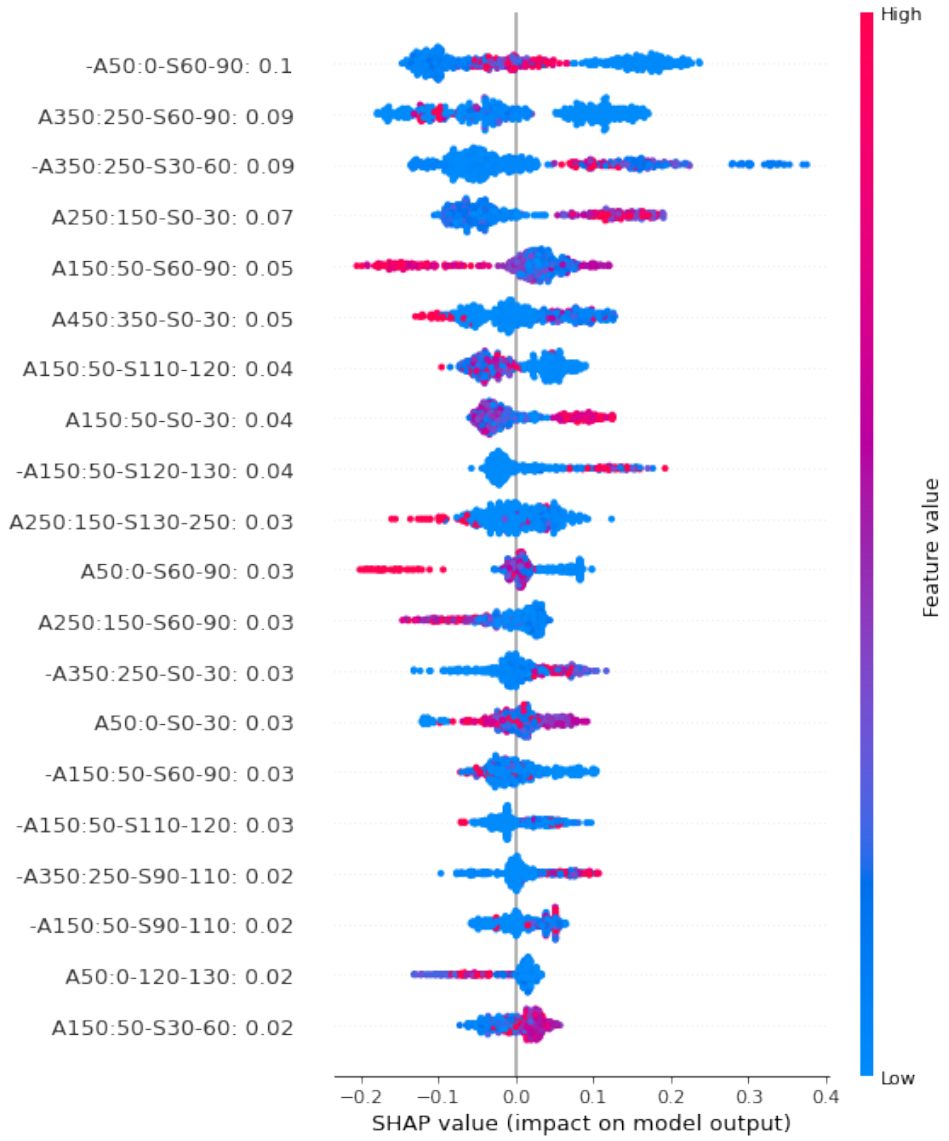

Fig. 5: SHAP Values Global Summary Ax

- 'A250:150-S0-30' signifies soft acceleration with a speed range of 0-30 km/h (SHAP = 7%).
- 'A150:50-S60-90' indicates soft acceleration with a speed range of 60-90 km/h (SHAP = 5%).
- 'A450:350-S0-30' represents soft acceleration with a speed range of 0-30 km/h (SHAP = 5%).
- 'A150:50-S120-130' signifies soft acceleration with a speed range of 120-130 km/h (SHAP = 4%).
- '-A150:50-S120-130' indicates soft braking with a speed range of 120-130 km/h (SHAP = 4%).
- 'A250:150-S130-250' corresponds to soft acceleration with a speed range of 130-250 km/h (SHAP = 3%).

For five of the top seven features, more in-depth interpretations can be drawn. Overlooking features '-A50:0-S60-90' and 'A450:350-S0-30' for which no clear trends emerge, we make the following interpretations for the remaining ones:

- **'A350:250-S60-90':** Drivers in the normal class tend to have high values in the speed segment S60:90.
- **'-A350:250-S30-60':** Drivers in the claims class tend to perform frequent soft braking in the speed zone S30-60. The frequency of braking is higher compared to drivers in the normal class.
- **'A250:150-S0-30':** Drivers in the claims class tend to have high values in bucket, indicating potential frequent traffic congestion. This is associated with low acceleration and weak braking.
- **'A150:50-S60-90':** Drivers frequently driving in the speed zone S60-90, with low , representing a low risk.
- **'A150:50-S110-120':** Drivers in the normal class are characterized by frequent driving in the speed zone S110-120, with low acceleration values.

IV. One class anomaly detection

In this section, we propose a new one-class supervised scheme for telematic-based risk assessment. This scheme does not require training on the underrepresented class (claims) as it treats claim detection as an anomaly detection problem. One class ML models seek identifying the anomalous data as those that lie on the decision boundaries inferred by training on the over-represented class only. Considering the one-class scheme offers key advantages from a modeling standpoint. First, as the claim data are very rare, the one-class model allows to put all of them in the test phase, making the results obtained much more representative than, say, if only 30% of the claim data were used for test and the remaining 70% for training. Second, it will give insights on how similar the behaviors of drivers without any claims are, and whether the behaviors of drivers with registered claims lie at the decision boundaries or not. In other words, this will allow us to investigate on whether drivers with claims show a driving behavioral pattern that is different from the pattern of drivers without claims. For this investigation, we consider two of the most popular one-class ML models, namely one-class SVM and Isolation forest.

A. One class SVM

One-Class SVM [34], a variant of Support Vector Machines (SVM), is used for outlier detection. It is an unsupervised model that seeks identifying the smallest hypersphere that comprises the bulk of data points while allowing some points, considered as outliers, to lie outside. The decision function for One-Class SVM is defined as: $f(x) = \text{sign}(w^T\phi(x) - b)$. The optimization problem can be solved using quadratic programming. Once the weight vector w and the bias term b are found, the decision function can be used to classify new data points as either inside the hypersphere or outside it.

B. Isolation Forest

Isolation Forest [28] is an unsupervised ML model used for anomaly detection. It is based on the concept of isolating anomalies (outliers) from normal data points by randomly partitioning through a binary tree structure. It relies on two main concepts: isolation and path length. Isolation is a measure of how easy it is to separate a data point from the rest of the dataset. Path length is a measure of how many partitions (or splits) it takes to isolate a data point. The lower the anomaly score, the more anomalous the data point is.

C. Experiments

To assess the performance of our one-class models, we use the whole claim data in the test set as none is considered in the training phase. The training and test sets regarding the normal (no claim) class remains the same as before. According to these settings, for the yearly data, the training and set sizes for class 0 (no claim) are 756 and 504 heatmaps while the test size for class 1 (claims) is 155. For the monthly data, these sizes are 5305, 3575 and 1084 respectively. For the weekly data, the sizes are 18890, 12725 and 3860 respectively. Tables **??** and **??** show the balance accuracy when testing one-class SVM and Isolation Forest respectively, on the weekly, monthly and yearly test sets. As observed, the best performance is obtained with Isolation Forest on the year data and reaches 53.2% of BA, which is barely above a random classifier achieving 50%. The results obtained are far below those obtained with the 2-class weighted XGBoost classifier. This implicitly reflects the fact that most claim data in the test set are not located at the boundaries of the normal data identified by one-class SVM or Isolation Forest. To confirm this intuition, we visualize, trough the t-SNE projection technique [36], the normal and predicted anomaly data according to one-SVM for the yearly, monthly and weekly data (Figures 6, 7, and 8 respectively). As observed, while the one-class SVM makes a good job as identifying outliers (figures in the middle), associating these outliers with claims is barely effective as the claims data strongly overlap with the normal data and barely lie at the boundaries (figures on the left). Even when we restrict the one-class SVM to identify only the 10% percentile of outliers, the latter barely overlap with the actual claims for the same reasons. The same observations can be drawn for the Isolation Forest algorithm as illustrated by Figures 9, 10, and 11 (for which the right figures have been dropped).

TABLE III: One-class SVM results for A_x speed

Data	Metric BA	Class	Precision	Recall	$F1$
week	0.515	class 0	0.77	0.99	0.87
		class 1	0.54	0.04	0.08
month	0.508	class 0	0.77	0.98	0.86
		class 1	0.38	0.03	0.06
year	0.517	class 0	0.77	0.98	0.86
		class 1	0.43	0.06	0.10

TABLE IV: Isolation Forest results for A_x speed

Data	Metric BA	Class	Precision	Recall	$F1$
week	0.517	class 0	0.77	0.93	0.85
		class 1	0.32	0.10	0.15
month	0.520	class 0	0.78	0.94	0.85
		class 1	0.34	0.10	0.16
year	0.532	class 0	0.78	0.96	0.86
		class 1	0.46	0.10	0.17

V. Drivers stability Assessment

The previous sections and the experiments therein have highlighted some issues with supervised 2-classes and 1-class models. The former assume a good separability between the two classes while the latter assume that claim data lie mainly at the boundaries of the normal data distribution. These assumptions, however, are barely supported by the results obtained: for the 2-classes setting, the results are promising given the heavily unbalanced class distribution but detecting a large proportion of claims is only obtained at the expense of misclassifying a large proportion of normal data. For the 1-class setting, the results show that the claim data barely

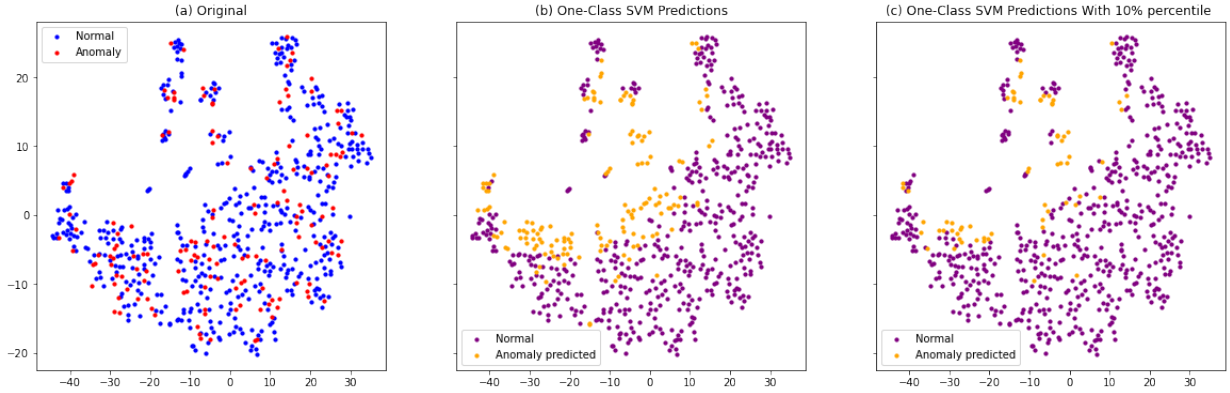

Fig. 6: Anomaly detection by the one-class SVM on the test at the **year** level: (a) actual distribution of normal (blue) and claim (red) data; (b) the distribution of normal (magenta) and claim (yellow) data as detected by the model; (c) the same distribution by retaining only 10% percentile of the anomaly data

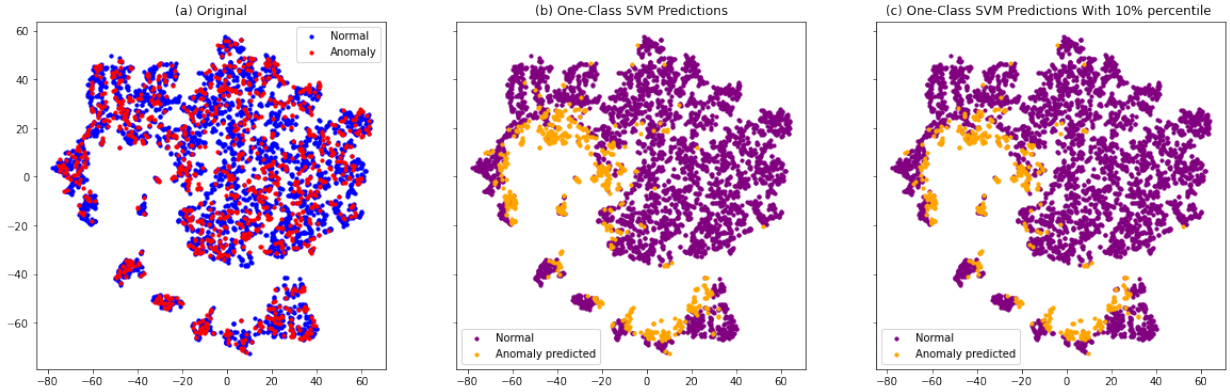

Fig. 7: Anomaly detection by the one-class SVM on the test at the **month** level: (a) actual distribution of normal (blue) and claim (red) data; (b) the distribution of normal (magenta) and claim (yellow) data as detected by the model; (c) the same distribution by retaining only 10% percentile of the anomaly data

lie at the boundaries as most show strong overlapping with the normal data. For these reasons, we propose a different modeling, inspired from works characterizing Alzheimer's disease from online handwriting [40] [23], that does not seek to discriminate claim from non-claim data, but rather detect homogeneous groups w.r.t behavior driving, and then analyse whether these groups are informative regarding claims or not. To this end, we propose a novel unsupervised ML technique consisting of a two-stage clustering scheme that not only allows to infer a driver signature and estimate, through sound information-theoretic measures, how stable a driver behavior is, but also uncovers the correlation between driver groups (clusters) and claim distributions. The first clustering stage uncovers behavioral clusters, i.e. monthly-based heatmap prototypes, irrespective of drivers' identities, while the second uncovers yearly-based driver clusters with similar behavior distribution over the first level clusters.

A. *First-Stage Clustering*

The first stage clustering seeks splitting the monthly aggregated telematic heatmaps into homogeneous groups (clusters) w.r.t driving behavior, irrespective of driver identity. This means that the driver monthly heatmaps may be assigned to different clusters or to the same cluster depending on their similarity in terms of driving behavior. To this end, we use the K-means algorithm to partition the monthly heatmaps into a predetermined number of clusters. The optimal number of clusters N_c obtained is equal to nine, determined based on the prediction strength technique [35]. Figure 12 shows the centroids of the nine clusters. This clustering is based on the $V - A_x$ distribution heatmap but clustering based on the other heatmaps can be performed in a similar way. As shown, each cluster represents a distinct pattern of driving behavior in terms of speed-acceleration joint distributions. The heatmaps within a same cluster are similar to the associated centroid and exhibit similar driving behavior over a month duration.

B. *Estimation of Driver (Un-)Stability Level*

The first-stage clustering uncovers heatmap clusters (groups) with each representing several drivers and each driver may have his/her monthly heatmaps assigned to several clusters. To infer a driver representation, we generate for each

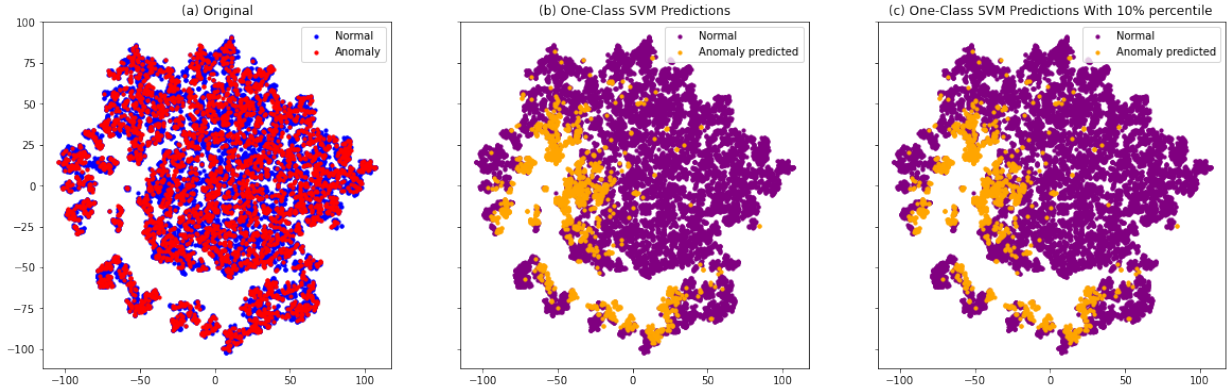

Fig. 8: Anomaly detection by the one-class SVM on the test at the **week** level: (a) actual distribution of normal (blue) and claim (red) data; (b) the distribution of normal (magenta) and claim (yellow) data as detected by the model; (c) the same distribution by retaining only 10% percentile of the anomaly data

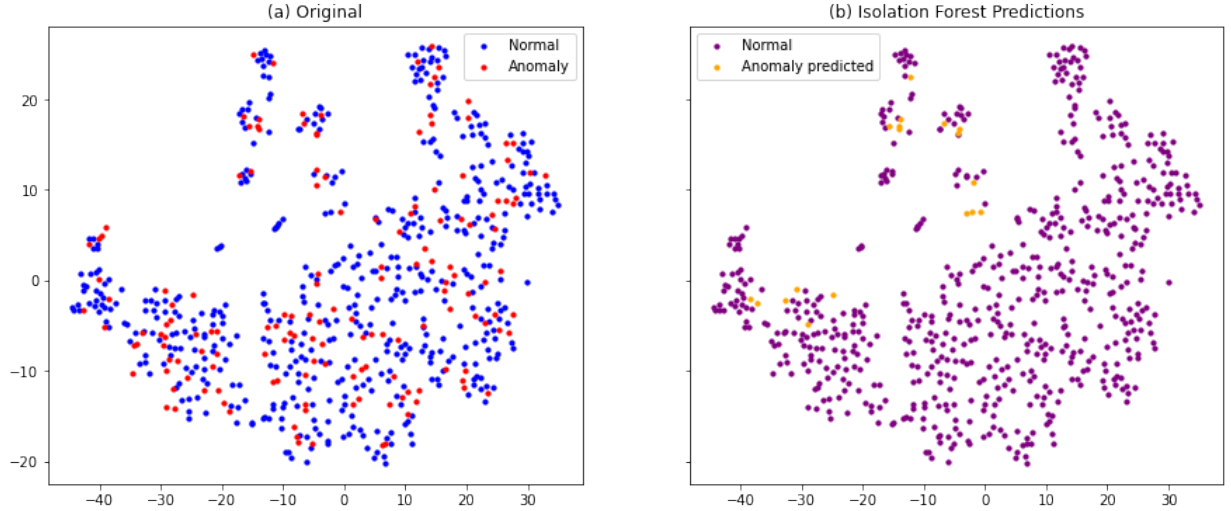

Fig. 9: Anomaly detection by Isolated Forest on the test at the **year** level: (a) actual distribution of normal (blue) and claim (red) data; (b) the distribution of normal (magenta) and claim (yellow) data as detected by the model

driver a histogram associated with the distribution of his/her monthly heatmaps over the different clusters (Figure 14). Owing to the variable number of months of driving experience for each driver, we normalize the histogram w.r.t this number to make the representation duration-independent. Each bin value, therefore, is between 0 and 1, and the bins' sum is equal to one. Intuitively, in the extreme case, if only one bin value is non-zero (equal to one) for a driver, this means that this driver produces roughly the same heatmap (driving behavior) at each month, implying thereby maximum stability. At the other extreme, if all the bins have the same frequency, i.e. $1/N_c$, this means that the driver produces all the heatmap clusters with equal probability, implying thereby a maximum instability. Between these two extremes, drivers show different levels of stability depending not only on how many of their heatmaps (bins) are activated but also on the frequencies (probabilities) of these activated heatmaps. To assess the driver instability level in a quantified way, we introduce information-theoretic measures based on entropy and perplexity. Entropy is defined as the degree of disorder, uncertainty or randomness, in a probabilistic system or source. We use it here to quantify the degree of instability of a driver in terms of the heatmaps he/she produces each month. The driver entropy is defined as:

$$H = -\sum_{i=k}^{N_c} p_k \log p_k \tag{3}$$

where log is the binary logarithm, and p_k is the probability of heatmap prototype (cluster) k, computed, in a maximum likelihood way, as its frequency over the driving experience duration. H reaches its minimum 0 for maximum driver stability (only one of the heatmap prototypes is activated) and its maximum $\log(N_c)$ at maximum instability, when all the prototypes are activated with equal probability $(1/N_c)$. Likewise, we introduce the driver perplexity, directly related to entropy through this equation: $P = 2^H$. Perplexity shows

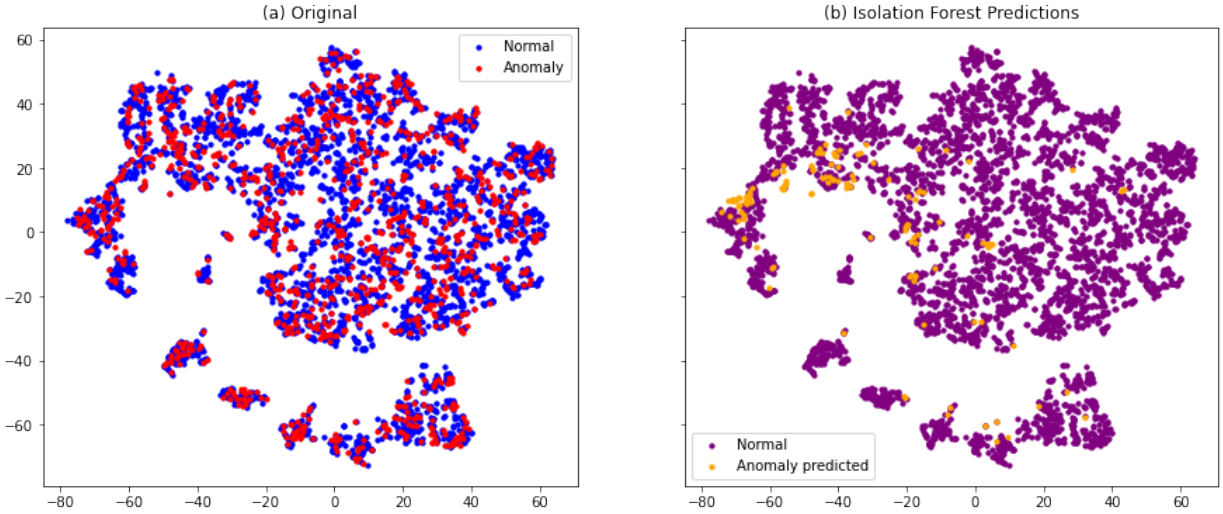

Fig. 10: Anomaly detection by Isolated Forest on the test at the **month** level: (a) actual distribution of normal (blue) and claim (red) data; (b) the distribution of normal (magenta) and claim (yellow) data as detected by the model

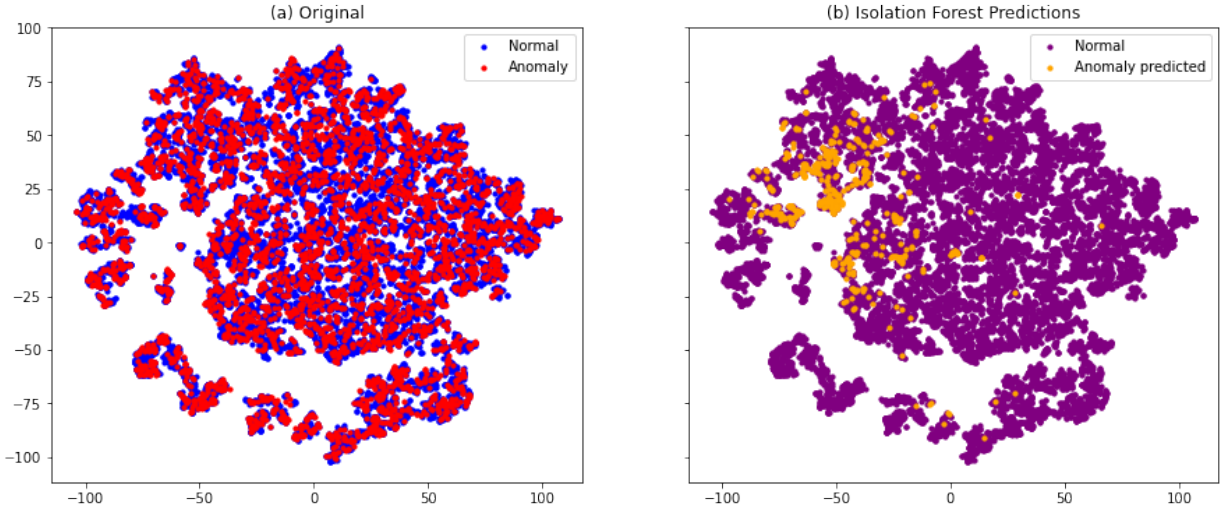

Fig. 11: Anomaly detection by Isolated Forest on the test at the **week** level: (a) actual distribution of normal (blue) and claim (red) data; (b) the distribution of normal (magenta) and claim (yellow) data as detected by the model

the same properties as entropy but it is easier to interpret as its minimum, 1, corresponds to the fact that only one heatmap prototype is activated, and its maximum, N_c, reflects the fact that all (the N_c) heatmap prototypes are activated with the same probability. Beyond these two extremes, the closer the perplexity to 1 the more stable the driver, and the closer the perplexity to N_c the more unstable the driver.

C. Second-Stage Clustering

Driver representation as a distribution over the 1-stage clusters has the additional key advantage that it allows inferring a fixed feature dimension (N_c) irrespective of the driver experience duration. As a result, we can now run any clustering algorithm on such representations to uncover now clusters (groups) of *drivers* with similar driving behavior, not only in terms of, speed-acceleration intervals' distribution, but also in terms of their stability/instability across their driving months. Based on the Kmeans algorithm and the prediction strength, we find an optimal number of clusters equal to 12 (shown in Figure 18). The clusters at the 2-stage clustering reflect similarity between drivers both in terms of low-level behavior (speed-acceleration intervals' distribution in the heatmaps) and high-level month-by-month behavior in terms of stability/instability across the monthly heatmap prototypes. This stability is shown by Figure 17 that displays the distributions of the 2-stage clusters over the heatmap prototypes (1-stage clusters). As illustrated, some clusters are much more stable than others. Based on the driver stability given by the perplexity computed previously, we can compute the stability level for each *cluster* as the average of the perplexities of the

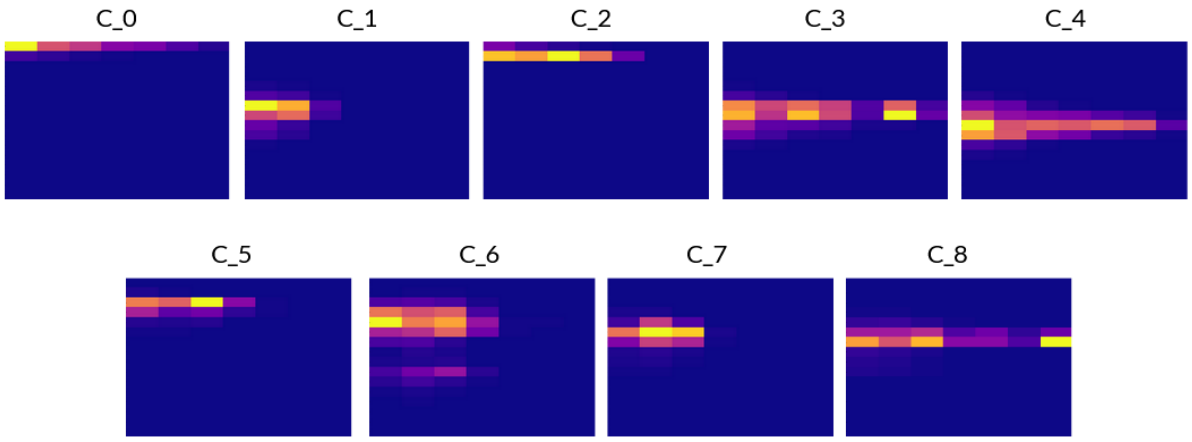

Fig. 12: The heatmaps' prototypes obtained at the first-stage clustering. For simplicity, the color scale and the speed / acceleration intervals are omitted; they are the same those in Figure 2

drivers assigned to that cluster. Table VII shows the average perplexities associated with the 2-stage clusters. Accordingly, the lowest the perplexity the more stable the drivers assigned to that cluster, and the higher the perplexity the more unstable the drivers in that cluster.

D. Correlation between claims and 2-stage clusters

Having performed 1-stage and 2-stage clustering, and provided an information-theoretic measure of the stability level of the 2-stage clusters, it remains to be seen whether there is correlation between the latter and the claim distribution. Table V shows the distribution of claims over the second-stage clusters. It shows the number of drivers in each cluster (N_i) and the number of claims for each cluster at different levels of responsibility (0_{resp}, 0.5_{resp}, 1_{resp}, 1.5_{resp}, 2_{resp}, 2.5_{resp}, and 3.5_{resp}). The last row shows the total number of claims in each cluster and the associated percentage. The last column shows the global claim distribution before clustering. In line with the results obtained with supervised modeling in Sections III and IV, no major trends are observed regarding the correlation between the clusters and the claims distributions. However, we do observe some partial correlations as the frequencies of the claims are higher in some clusters w.r.t the global distribution before clustering, and lower in some others. Intuitively, more informative clusters are those that have the most different claim distributions w.r.t the global one as they correspond to significantly either higher or lower claims' occurrence w.r.t the global initial distribution before clustering. Specifically, in Table V, apart from the natural fact that each cluster is dominated by the 0 claim frequency giving that this class has a large global frequency of 91.18%, we observe that the global claim percentage is 8.72%, while this percentage decreases significantly to 5.3% and 5.71% for clusters 3 and 5 respectively. Conversely, clusters 4 and 7 have significantly more claims w.r.t the global claim distribution with ratios of 12.6% and 12.38%. We may partially interpret this finding by arguing that the first two clusters are associated with more risky driving behavior and that the drivers with no claim inside may be seen as those with risky behavior although they were not involved in accidents. The two last clusters, by contrast, are characterized by less risky behavior and the drivers with claims inside may be seen as those with less risky behavior although they happened to be involved in accidents.

The interpretations above are based only on the global clusters' claim distributions without taking into account the fine levels of claim responsibility. To tackle this point and assess cluster correlation level with claim distribution in a fine way, we introduce the Kullback–Leibler (KL) clusters' divergence between each cluster and the global one (the one with no clustering, comprising all drivers). Kullback–Leibler divergence is a distance metric that quantifies the difference between two probability distributions. Before we apply the KL divergence, we convert the cluster claim distributions in Table V into probabilities by normalizing each value column-wise and smoothing them. We obtain in this way the claim probability distribution for each cluster and also the global one before clustering. Formally, let P be the global label probability distribution of classes (claims), and Q_k the label probability distribution knowing cluster C_k ($k = 1...N_c$). The KL divergence is then calculated as follows:

$$KL(P \parallel Q_k) = D(P \parallel Q_{C_k}) = \sum_{i=0}^{N_L-1} P(i) \log \frac{P(i)}{Q_k(i)} \quad (4)$$

where N_L is the number of classes or labels (N_L = 7), $i = 0; 1; ..., 6$ is associated with levels of claims responsibility equal to 0; 0.5; 1; 1.5; 2; 2.5; 3.5 respectively, and $P(i)$ and $Q_k(i)$ are the probabilities of class i under P and Q_k, computed in a maximum likelihood way, as follows:

$$P(i) = \frac{N_i}{N} \quad (5)$$

$$Q_k(i) = Prob_{Q_k}(i/C_k) = \frac{N_{ik}}{N_k} \quad (6)$$

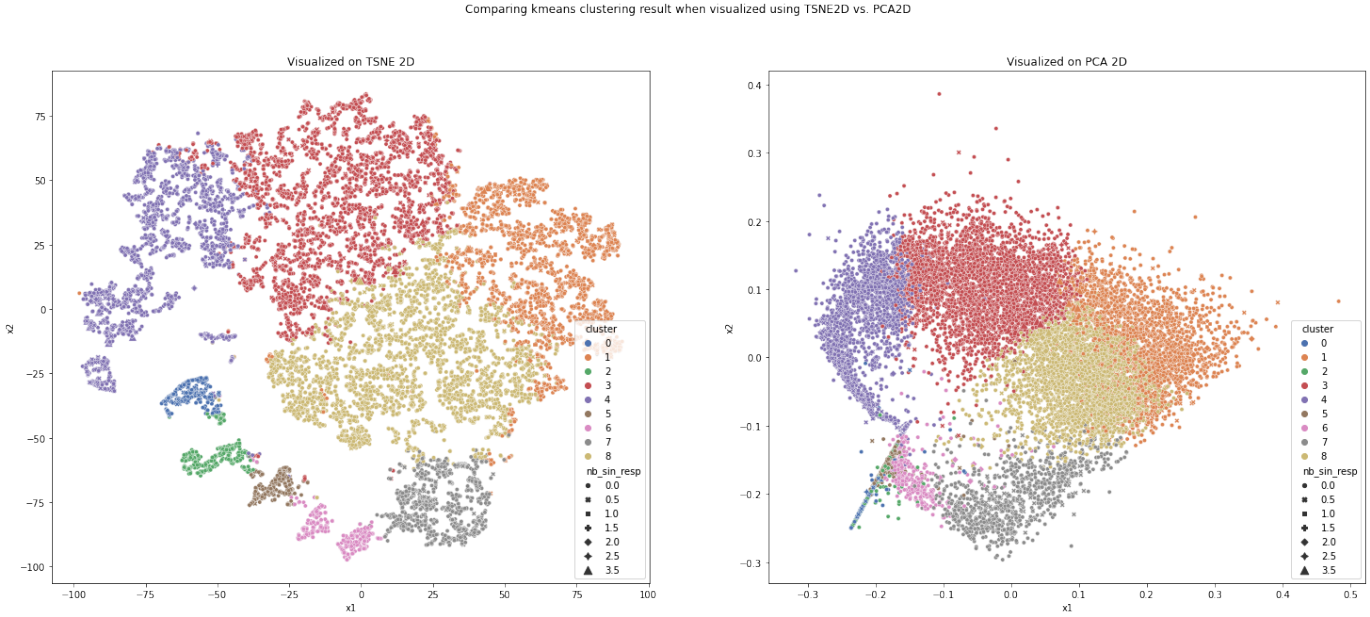

Fig. 13: Results of k-means first clustering with the optimal number, nine k=9, is the maximum number of clusters with a prediction strength ≥ 0.8

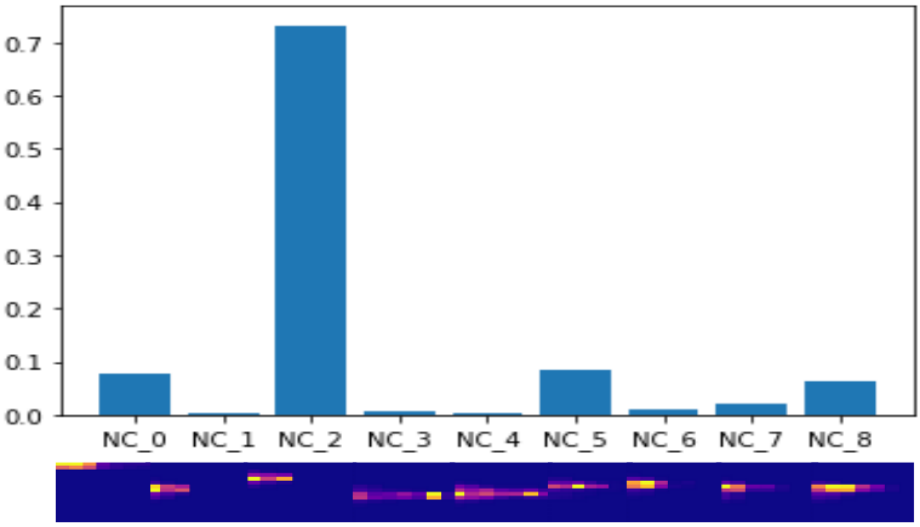

Fig. 14: Distribution of one driver monthly telematic data over the heatmaps' prototypes obtained at the first-stage clustering

where N is the total number of drivers, N_i is the number of drivers with claim label of index i, $k = 1, ..., N_{C2}$, with N_{C2} the number of clusters at the second stage, N_k the number of drivers in cluster C_k, and N_{ik} the number of drivers with claim label of index i in cluster C_k. We also have $N = \sum_i^{N_L} N_i$. Based on these definitions, we compute the global KL divergence between the global distribution and the whole clustering scheme as the weighted sum of the cluster-based KL divergences.

$$D(P \parallel Q) = \sum_{k=1}^{C_k} D(P \parallel Q_{C_k}) W_k \tag{7}$$

where $W_k = N_k/N$ is the percentage of heatmaps pertaining to cluster C_k, used as a coefficient to give more weight to denser clusters.

Table VI shows the KL divergences between each (2-stage) cluster distribution and the global distribution, with the no-claim class considered (first row) and without (second row). As observed in the first row, no clusters having a really dissimilar distribution from the global one actually emerge, which can be explained by the heavy dominance of the no-claim class over all the distributions. By taking out this class, and normalizing the remaining probabilities accordingly, clusters C_7 and C_8 clearly emerge with a KL much more higher than all the

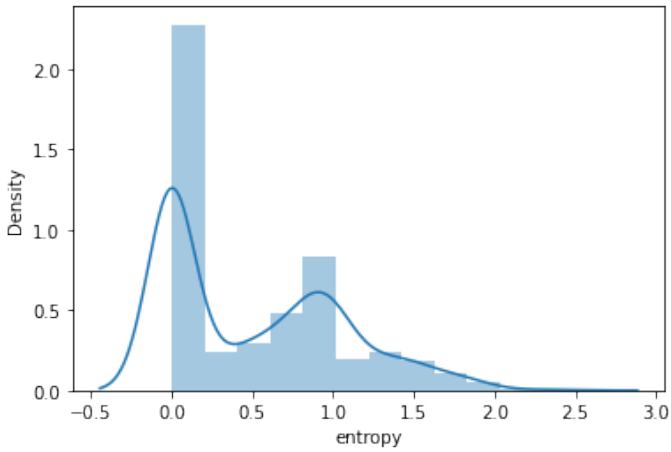

Fig. 15: Entropy

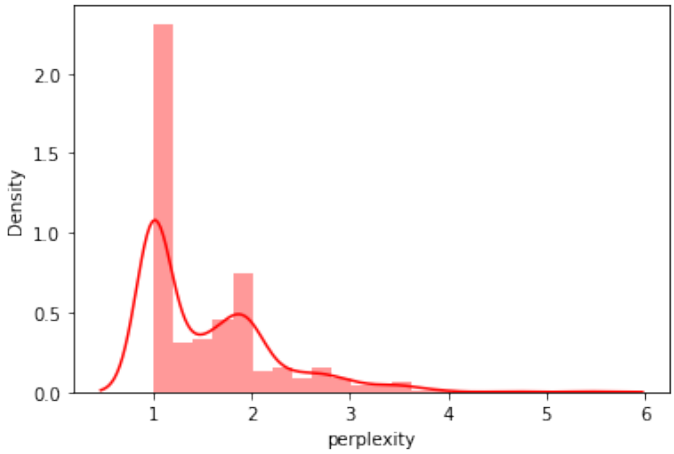

Fig. 16: Perplexity

others. This is understandable, as these two clusters do not comprise any claim data (Table V). These two clusters may be associated with very safe driving. The third most different cluster with the global distribution is C_5 that comprises only one driver with claims. This driver may be associated with safe driving although he/she has been involved in a claim. Another interesting cluster is C_{10} with a fair KL. It has a much larger relative number of claims w.r.t the global distribution. The drivers with no-claims therein my be seen as drivers with risky behavior even though they happen not to have been involved with claims. These results show the ability of our 2-stage clustering scheme to uncover different clusters of drivers with behaviors associated with different driving risks and claim distributions.

Globally, the obtained KL divergence between the whole 2-stage clustering ($N_{C2} = 12$) and the global distribution is D = 0.023, while the same divergence without taking into account the no-claim class increases to 0.078. These two values can be used to compare different clustering methods in order to identify the one providing the best correlation with claims.

VI. Conclusion

The objective of this paper was to investigate whether driver behavior, as reflected by in-car collected telematic data, is correlated with risks leading to involvement in claims. Our extensive investigation based on supervised and unsupervised models demonstrates that, although the best balanced accuracy obtained, i.e. 62% is clearly far below a satisfactory performance that can be harnessed for insurance pricing, the answer is yes as such correlation, albeit partially, does exist. This is reflected by the significant increase by 12% of balanced accuracy w.r.t a blind classifier and the correlation that exists between some clusters and claims, despite the adverse conditions in which our models have operated. These conditions are severalfold. First, the claim's frequency, even when all the levels of responsibility are merged, is rare as it consists of less than 9% of the data. This unbalanced issue is further worsened by the fact that they are split into Training, validation and test data. As a matter of fact, this rarity implies that the distributions of the telematics data distributions associated with claims over these subsets consisting of different drivers may show low overlapping, making optimization of models parameters based on the first two hardly generalizable to the unseen test data. Second, most of the claims have occurred years before the experimental driving period, which may decrease significantly the correlation between the claims and the collected telematic data. Third, the telematic-based feature heatmaps are extracted without taking into account key contextual information such as the speed limit in the driving zone, which hampers drastically the quality of the features extracted. As an example, a speed of 120% km/h is normal in a highway with 130% as speed limit and extremely dangerous in an extra urban zone with 90 km/h as speed limit. Other important contextual indication such as night/day period and weather conditions, even though if they are available, have been dropped owing to the rarity of the claim distribution, making the addition of context data likely to worsen the generalization capabilities of our models.

To significantly improve performance, a natural direction is to tackle the issues mentioned above. Regarding the heavy class unbalance problem reflected by the rarity of the claim class, our extensive experiments with different models shows that additional investigations of other models is likely to be doomed to failure, as the main issue is not model-related but data-related. Rather than following a model-based approach, we should, therefore, adopt a data-centric approach, a recent paradigm-shift trend in machine learning that has gained attention over the last couple of years [32]. Data-centric AI trains the model and hyperparameters once and keep it fixed. Accuracy is improved by data-driven error analysis, in which model errors are analysed in terms of the worst-recognised classes and the main types of mis-classified inputs. This guides a smart collection of new data to minimise such errors, which is much more effective than randomly collecting data for standard model-centric approaches. Once sufficient training claim data are collected in this way to cover fairly enough different feature-context configurations, contextual information

TABLE V: Distribution of claims over the 2-stage clusters, for A_x

claims	$cluster_0$	$cluster_1$	$cluster_2$	$cluster_3$	$cluster_4$	$cluster_5$	$cluster_6$	$cluster_7$	$cluster_8$	$cluster_9$	$cluster_{10}$	$cluster_{11}$
0_{sinis}	71	154	149	52	51	17	107	20	20	24	90	187
0.5_{resp}	7	9	10	3	6	1	9	0	0	2	11	11
1_{resp}	0	1	0	0	0	0	2	0	0	0	1	0
1.5_{resp}	0	0	2	0	0	0	0	0	0	0	0	0
2_{resp}	0	1	1	0	0	0	2	0	0	1	2	1
2.5_{resp}	0	0	1	1	1	0	0	0	0	0	0	2
3.5_{resp}	1	0	1	0	0	0	0	0	0	0	0	0
groups	8	11	15	4	7	1	13	0	0	3	14	14
%	10.12%	6.66%	9.14%	7.27%	12.06%	5.55%	10.83%	0%	0%	11.11%	13.46 %	6.97%

TABLE VI: KL second clustering k=12 A_x

Cluster	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
D_{pq}second clustering k=12	0.01	0.021	0.01	0.009	0.019	0.095	0.01	0.088	0.088	0.073	0.017	0.029
D_{pq} second clustering only claims k=12	0.04	0.035	0.051	0.057	0.03	0.134	0.063	0.416	0.416	0.094	0.06	0.065

TABLE VII: Second clustering perplexity k=12

Cluster	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}
mean perplexity	2.272	1.288	1.218	2.553	1.497	1.306	1.521	1.749	1.828	1.418	2.295	1.081
std perplexity	0.583	0.345	0.485	0.555	0.674	0.597	0.392	1.023	1.017	0.685	0.587	0.184

and speed limit that might be collected from GPS devices should be added to the telematic data, to make the latter more informative regarding driver behavior. An other improvement direction is data augmentation that consists of synthesizing under-represented data, i.e. the claim data. Our first endeavors in this direction based on the SMOTE technique were not successful. We argue that Generative Adversarial Networks (GAN)-based augmentation techniques can be more effective for data augmentation, whether for image-like input [19][20], or telematics data stream time-series [8], especially by adapting our recent work that selects the synthesized data with the most impact on reducing errors [26]. Finally, authenticating car drivers should also be included to disentangle the telematic data collected from different drivers using the same car.

REFERENCES

[1] Subramanian Arumugam and R Bhargavi. "A survey on driving behavior analysis in usage based insurance using big data". In: *Journal of Big Data* 6.1 (2019), pp. 1–21.

[2] Mercedes Ayuso, Montserrat Guillen, and Ana María Pérez Marín. "Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance". In: *Transportation research part C: emerging technologies* 68 (2016), pp. 160–167.

[3] Mercedes Ayuso, Montserrat Guillen, and Jens Perch Nielsen. "Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data". In: *Transportation* 46 (2019), pp. 735–752.

[4] Mercedes Ayuso, Montserrat Guillen, and Ana María Pérez-Marín. "Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's". In: *Risks* 4.2 (2016), p. 10.

[5] Christopher Blier-Wong et al. "Machine learning in P&C insurance: A review for pricing and reserving". In: *Risks* 9.1 (2020), p. 4.

[6] Jean-Philippe Boucher, Steven Côté, and Montserrat Guillen. "Exposure as duration and distance in telematics motor insurance using generalized additive models". In: *Risks* 5.4 (2017), p. 54.

[7] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2016.

[8] Quang Dao, Mounim A. El-Yacoubi, and Anne-Sophie Rigaud. "Detection of Alzheimer Disease on Online Handwriting Using 1D Convolutional Neural Network". In: *IEEE Access* 11 (2023), pp. 2148–2155. DOI: 10.1109/access.2022.3232396.

[9] European Data Protection Board. *EDPB Guidelines 202001 Connected Vehicles*. Available at: https://edpb.europa.eu/sites/default/files/consultation/edpb_guidelines_202001_connectedvehicles.pdf. 2020.

[10] Jingzhong Fang et al. "A Survey of Algorithms, Applications and Trends for Particle Swarm Optimization". In: *International Journal of Network Dynamics and Intelligence* (Feb. 2023).

[11] Guangyuan Gao, Shengwang Meng, and Mario V Wüthrich. "Claims frequency modeling using telematics car driving data". In: *Scandinavian Actuarial Journal* 2019.2 (2019), pp. 143–162.

[12] Guangyuan Gao, He Wang, and Mario V Wüthrich. "Boosting Poisson regression models with telematics car driving data". In: *Machine Learning* 111.1 (2022), pp. 243–272.

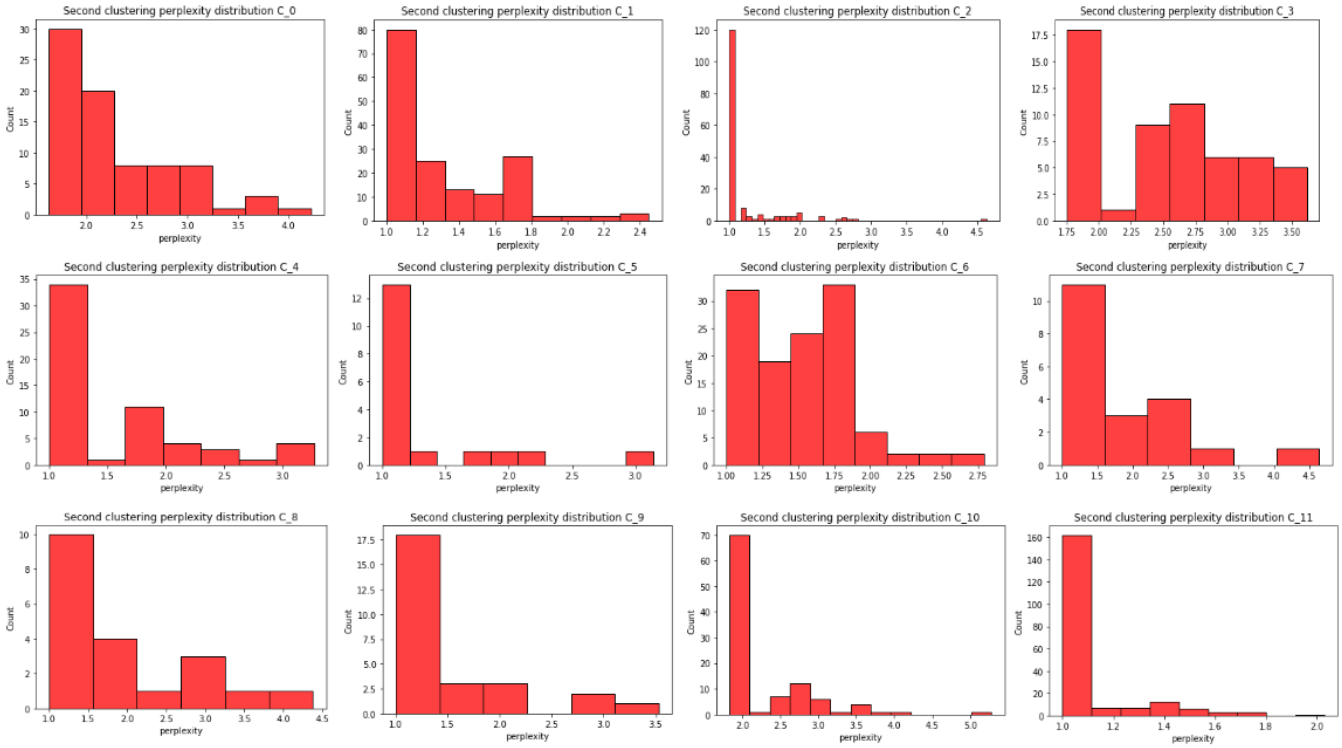

Fig. 17: Perplexity distribution over the 2-stage clusters

[13] Guangyuan Gao and Mario V Wüthrich. "Convolutional neural network classification of telematics car driving data". In: *Risks* 7.1 (2019), p. 6.

[14] Guangyuan Gao and Mario V Wüthrich. "Feature extraction from telematics car driving heatmaps". In: *European Actuarial Journal* 8.2 (2018), pp. 383–406.

[15] Guangyuan Gao, Mario V Wüthrich, and Hanfang Yang. "Evaluation of driving risk at different speeds". In: *Insurance: Mathematics and Economics* 88 (2019), pp. 108–119.

[16] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. "Why do tree-based models still outperform deep learning on tabular data?" In: (2022).

[17] Montserrat Guillen, Jens Perch Nielsen, and Ana M Pérez-Marín. "Near-miss telematics in motor insurance". In: *Journal of Risk and Insurance* 88.3 (2021), pp. 569–589.

[18] Montserrat Guillen et al. "Can automobile insurance telematics predict the risk of near-miss events?" In: *North American Actuarial Journal* 24.1 (2020), pp. 141–152.

[19] Hajar Hammouch et al. "A two-stage Deep Convolutional Generative Adversarial Network-based data augmentation scheme for agriculture image regression tasks". In: *2021 International Conference on Cyber-Physical Social Intelligence (ICCSI)*. IEEE, Dec. 2021. DOI: 10.1109/iccsi53130.2021.9736230.

[20] Hajar Hammouch et al. "GANSet - Generating annnotated datasets using Generative Adversarial Networks". In: *2022 International Conference on Cyber-Physical Social Intelligence (ICCSI)*. IEEE, Nov. 2022. DOI: 10.1109/iccsi55536.2022.9970561.

[21] Mohamed Hanafy and Ruixing Ming. "Machine learning approaches for auto insurance big data". In: *Risks* 9.2 (2021), p. 42.

[22] Roel Henckaerts et al. "Boosting insights in insurance tariff plans with tree-based machine learning methods". In: *North American Actuarial Journal* 25.2 (2021), pp. 255–285.

[23] Christian Kahindo et al. "Characterizing early-stage Alzheimer through spatiotemporal dynamics of handwriting". In: *IEEE Signal Processing Letters* 25.8 (2018), pp. 1136–1140. DOI: 10.1109/lsp.2018.2794500.

[24] Mohamed Ibn Khedher, Mounim A. El-Yacoubi, and Bernadette Dorizzi. "Fusion of appearance and motion-based sparse representations for multi-shot person re-identification". In: *Neurocomputing* 248 (July 2017), pp. 94–104. DOI: 10.1016/j.neucom.2016.11.073.

[25] Jean Lemaire, Sojung Carol Park, and Kili C. Wang. "THE USE OF ANNUAL MILEAGE AS A RATING VARIABLE". In: *ASTIN Bulletin: The Journal of the IAA* 46.1 (2016), pp. 39–69.

[26] Yantao Li et al. "Adaptive Deep Feature Fusion for Continuous Authentication with Data Augmentation".

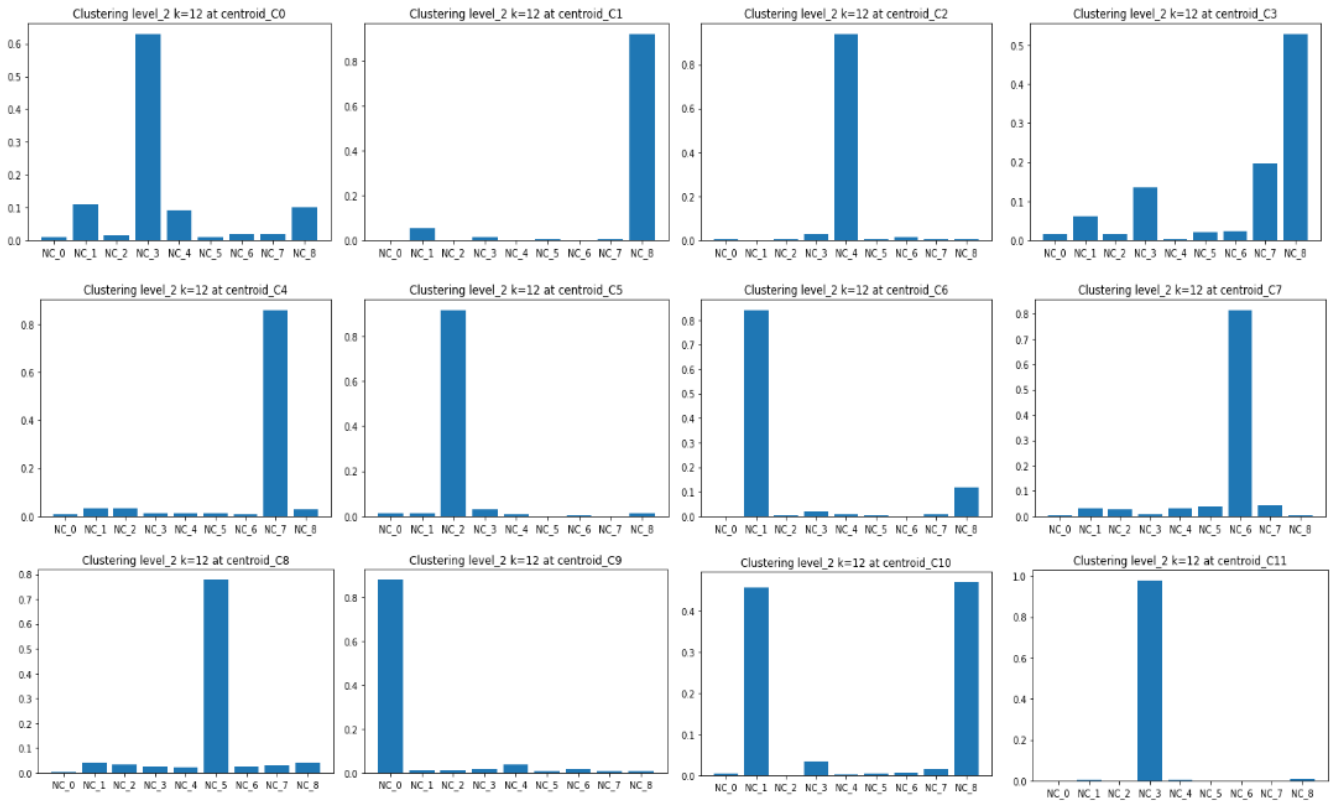

Fig. 18: Distribution of the 2-stage clusters over the heatmap prototypes (1-stage clusters)

In: *IEEE Transactions on Mobile Computing* (2022), pp. 1–16. DOI: 10.1109/tmc.2022.3186614.

[27] Todd Litman. "Pay-As-You-Drive Pricing and Insurance Regulatory Objectives." In: *Journal of Insurance Regulation* 23.3 (2005).

[28] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation Forest". In: *2008 Eighth IEEE International Conference on Data Mining*. 2008, pp. 413–422.

[29] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems* 30 (2017), pp. 4765–4774.

[30] Huiying Mao et al. "Decision-adjusted driver risk predictive models using kinematics information". In: *Accident Analysis & Prevention* 156 (2021), p. 106088.

[31] Johannes Paefgen, Thorsten Staake, and Elgar Fleisch. "Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data". In: *Transportation Research Part A: Policy and Practice* 61 (2014), pp. 27–40.

[32] Hima Patel et al. "Advances in Exploratory Data Analysis, Visualisation and Quality for Data Centric AI Systems". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2022.

[33] Jessica Pesantez-Narvaez, Montserrat Guillen, and Manuela Alcañiz. "Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression". In: *Risks* 7.2 (2019). ISSN: 2227-9091. DOI: 10.3390/risks7020070.

[34] Bernhard Schölkopf et al. "Estimating the support of a high-dimensional distribution". In: *Neural computation* 13.7 (2001), pp. 1443–1471.

[35] Robert Tibshirani and Guenther Walther. "Cluster Validation by Prediction Strength". In: *Journal of Computational and Graphical Statistics* 14.3 (Sept. 2005), pp. 511–528.

[36] L.J.P. van der Maaten and G.E. Hinton. "Visualizing High-Dimensional Data Using t-SNE". English. In: *Journal of Machine Learning Research* 9.nov (2008). Pagination: 27, pp. 2579–2605. ISSN: 1532-4435.

[37] Roel Verbelen, Katrien Antonio, and Gerda Claeskens. "Unravelling the predictive power of telematics data in car insurance pricing". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67.5 (2018), pp. 1275–1304.

[38] Mario V Wüthrich. "Covariate selection from telematics car driving data". In: *European Actuarial Journal* 7.1 (2017), pp. 89–108.

[39] Mario V Wüthrich and Michael Merz. *Statistical foundations of actuarial learning and its applications*. 2023.

[40] Mounım A. El-Yacoubi et al. “From aging to early-stage Alzheimer’s: Uncovering handwriting multimodal behaviors by semi-supervised learning and sequential representation learning”. In: *Pattern Recognition* 86 (Feb. 2019), pp. 112–133. DOI: 10.1016/j.patcog.2018.07.029.

[41] Rui Zhu and Mario V Wüthrich. “Clustering driving styles via image processing”. In: *Annals of Actuarial Science* 15.2 (2021), pp. 276–290.

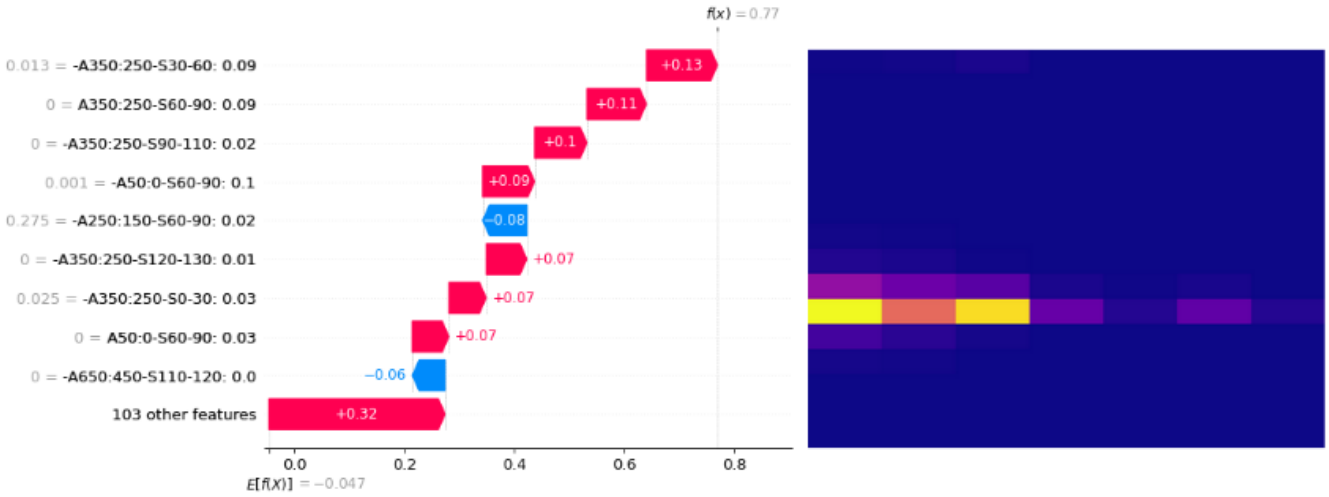

Fig. 19: TP Driver_id=051197, prob=0.6835044