European Study Group with Industry

ESGI 151 - Team No. 3

HOLON TECHNOLOGIES:
Usage based insurance

Authors:

Dusan Bikov (MACEDONIA)
Raul Kangro (ESTONIA)
Harald Kitzmann (GERMANY)
Petre Lameski (MACEDONIA)
Alice Raffaele (ITALY)
Zlatko Varbanov (BULGARIA)

Company members:

Raul Orav (ESTONIA)
Inga Murula (ESTONIA)
Dan Sööl (ESTONIA)

University of Tartu (ESTONIA),
February 4-8, 2019

Contents

List of Figures

1 Introduction

The *First Estonian Study Group with Industry* [3] was held at the Mathematics and Computer Science Department in the University of Tartu, with about fourty researchers, PhD students and professors from several countries in Europe and other foreign states. During this event, four companies proposed problems they are dealing with, asking for consulting and support to solve them.

1.1 Holon Technologies

Holon Technologies [7], one of the companies involved, works in the insurance sector, exploiting telematics and GPS data in order to provide useful indications to insurance companies which offer usage-based policies. These provide people with the same coverage that conventional auto insurance policies do, but are often less expensive (see Fig. 1).

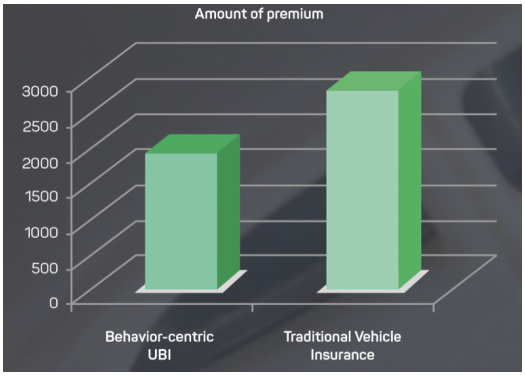

Figure 1: Usage based insurance vs Traditional insurance

When safe driving habits are demonstrated through telematics data, customers could save quite a lot. In fact, customers receive benefits such as reduced invoices, discounts, driving feedback; most importantly, the pricing is fair and is consistent with the customer's driving behavior, which is compared to the average driving style (see Fig. 2).

Holon Technologies, as a business partner, binds together different partner interests and provides know-how and technology that enable to offer efficient

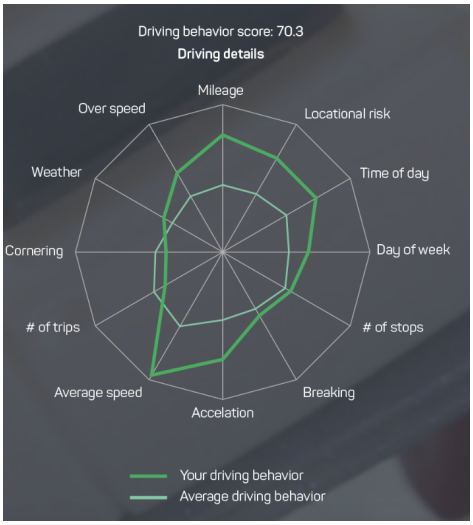

Figure 2: Driving details

products and services to final customers, which could be insurance companies or logistics ones, in charge of fleet management.

Its mission is to provide clients efficient solutions for data aggregation and analytics. Their vision is to become the number one platform for volume sensitive insurance service provision in Europe.

1.2 Usage based models

In the latest years, the development and spread of big data and Internet-of-Things technologies have enabled to start collecting a huge quantity of data about pretty much everything.

For instance, consider the data collected when surfing the net looking for some items to buy: customers searches, preferences and buyings are registered and often e-commerce websites propose other desirable objects similar to the previous already purchased; beyond these suggestions, there is an analysis of customers behaviour during their navigation.

Deepening the insurance sector, traditional insurances which consider mainly the car type and usage are being overcome by new usage based model, that calculate the premium according to data collectable: driving behaviors, trips, claims and crashes details, premium, etc..

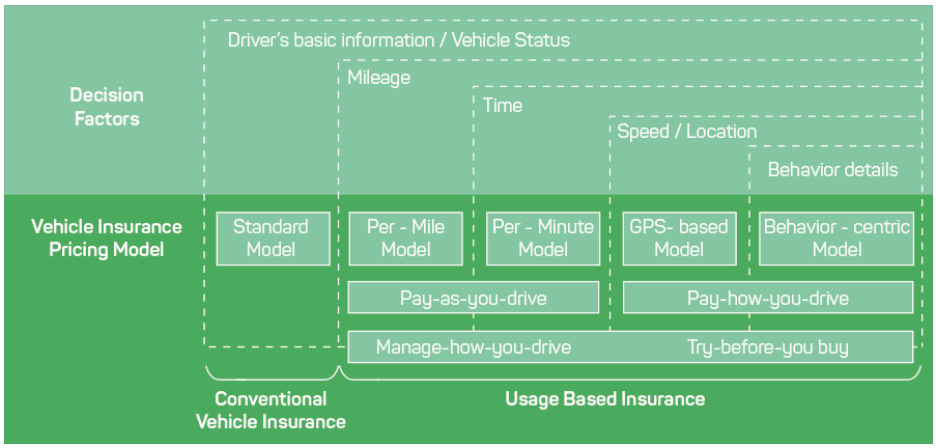

Figure 3: Usage based insurance models

There exist four main types of usage based models:

- **Pay-As-You-Drive (PAYD)**: Insurance premium is computed dynamically according to the amount driven (e.g., based on the odometer, on mileage aggregated from GPS data or on driving time);
- **Pay-How-You-Drive (PHYD)**: Similar to PAYD, it also uses additional sensors (e.g., accelerometer);
- **Manage-How-You-Drive (MHYD)**: Drivers receive periodic report about their driving style and they can adjust it to avoid the increase of premium;
- **Try-Before-You-Buy (TBYB)**: Insurers offer an app for free to potential customers, to analyze their driving style, promising a discount on a new policy if their score is good.

1.3 The problem

In order to become a leader in providing these kind of models to clients, Holon Technologies needs to develop efficient methods how to use driving behavior data, combined with accidents information, to make sensible predictions about possible damages.

Several heterogeneous sources must be taken into consideration: GPS data, telematics, traffic and roads information, speed limit, past accidents and crashes, etc.

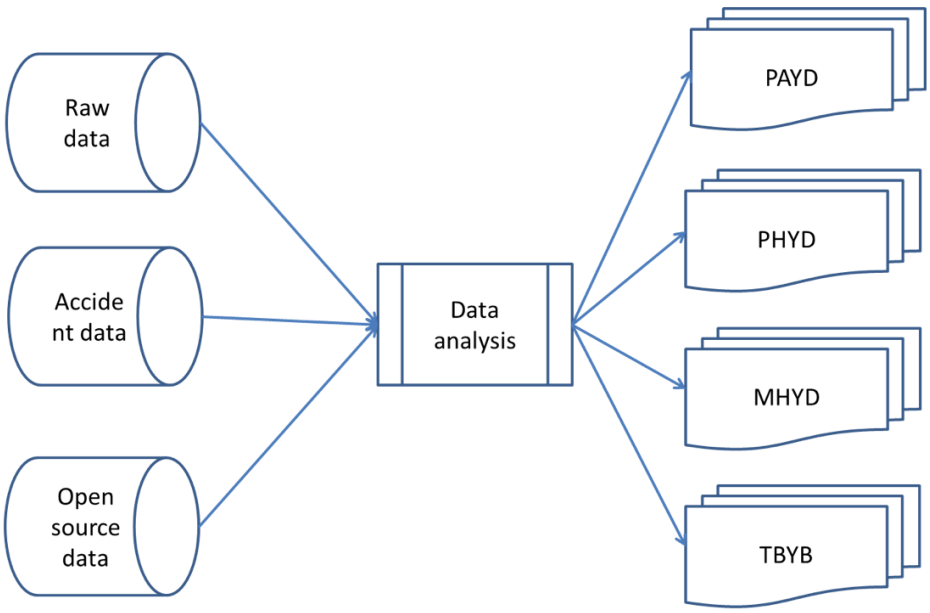

Figure 4: Decision process

Main questions asked from Holon Technologies to this study group are the following:

1. *Which method and techniques could be used to classify the data?*
2. *How to find a (reasonable) trend for drivers' behavior and how to compute the probability of the trend to continue over time?*
3. *How to define a model, combining different aspects, to determine the probability of damage, the damage frequency and the probable amount of the damage? Therefore how the premium should be changed? And what about pricing?*

1.4 Report structure

The report is organized as follows. Chapter 2 contains a brief state of the art about usage based insurance and existing methods developed. Holon Technologies provided some data extracted from their databases: they are described and, after brainstorming, analyzed in Chapter 3 and Chapter 4 respectively. Our proposed methodologies and techniques to tackle the problem are described in Chapter 5. Finally conclusions are in Chapter 6.

2 State of the Art

Usage-based insurance has become a topic of interest in last ten years, after the developing and spreading of big data and Internet-of-Things technologies, together with machine learning techniques.

The idea of modelling high-dimensional insurance data can be already found in works of 2004, such as Christmann [1], where some common features in data sets from motor vehicle insurance companies are described. There is proposed a general approach in order to model them with a complex dependency structure, using kernel logistic regression and ϵ-support vector regression. Later, these ideas are implemented in another joint work of Christmann and Galiano [8].

More recently, in 2017 Gao and Wuthrich [9, 5] analyzed and classified high-frequency GPS location data of several car drivers, introducing a so called *v-a heatmap*, which displays the average speed on the x-axis (in km/h) and the corresponding acceleration/braking pattern on the y-axis (in m/s^2). They differentiated various driving styles using *K-means* algorithm. Moreover, they applied singular value decomposition and bottleneck neural networks for Principal Component Analysis.

Gao, Meng and Wuthrich [4] analyzed 1.2 TB of telematics data and extracted covariates suitable for car insurance pricing.

At the beginning of 2019, Gao and Wuthrich [6] exploited convolutional neural networks to predict individual driving style in the future, extracting feature information about driving style from high-frequency GPS location data of individual car drivers and trips, constructing time series.

Despite these results, we underline that, as far as we know, a standard model, which takes into account both claims and driving behaviors, is still absent.

3 Data Description

Three types of input data could be identified which are divided by the types of sources. Raw data are data given as telematics and stored in each vehicle and describes the individual behaviour of each vehicle and its drivers during the lifetime and usage of the vehicle. Accident data are provided by the insurance organisations and describes the characteristics of each single accident (insurance case). The third group of data provided by state and other organisation as public available data like weather data, road conditions, etc.

3.1 Available data

This chapter describes main kinds of information provided by the company, stored in a Microsoft SQL Server and reachable through Microsoft Remote Desktop Connection. The three most important tables in the database are *Raw data*, *Trips data* and *Crash data*.

	timestamp	timestampUTC	EngineStatus	MovementSensor	GreenDrivingValue	GreenDrivingType	SplitSegment
0	2017-05-27 04:05:08.000	300	1.0	NaN	NaN	NaN	NaN
1	2017-05-27 04:06:08.000	300	1.0	NaN	NaN	NaN	NaN
2	2017-05-27 04:06:18.000	300	1.0	NaN	NaN	NaN	NaN
3	2017-05-27 04:07:08.000	300	1.0	NaN	NaN	NaN	NaN
4	2017-05-27 04:07:20.000	300	1.0	NaN	NaN	NaN	NaN

Distance	Direction	Power	Longitude	Latitude	GPSState	DriverID	Speed	DeltaDistance	Fuel	vehicleID
0.0	126.2	14.003	25.567075	58.892853	1.0	NaN	0.0	NaN	NaN	563
0.0	44.7	13.969	25.567090	58.892858	1.0	NaN	0.0	NaN	NaN	563
0.0	44.7	14.001	25.567088	58.892860	1.0	NaN	0.0	NaN	NaN	563
0.0	44.7	14.025	25.567088	58.892860	1.0	NaN	0.0	NaN	NaN	563
0.0	44.7	14.008	25.567088	58.892860	1.0	NaN	0.0	NaN	NaN	563

Figure 5: Raw data

3.1.1 Raw data

This table contains GPS points information for each vehicle collected during driving time, such as movements from sensors and speeds; 865,227 rows of different raw data were available in an extract of the whole database.

3.1.2 Trips data

This table shows vehicle aggregated data focusing on trips, based on the Raw data; 2,438,175 rows were available in the database example.

	vehicleID	avgSpeed	duration	distance	stoppedAfter	maxSpeed	startTimestamp	endTimestamp
0	361	54.0	1336.0	19.92	1771.0	97.0	2016-01-19 07:32:38.000	2016-01-19 07:54:53.000
1	361	41.0	1510.0	17.23	14142.0	91.0	2016-01-19 08:24:24.000	2016-01-19 08:49:33.000
2	15	0.0	12.0	0.07	1064.0	14.0	2016-01-19 10:54:58.000	2016-01-19 10:55:09.000
3	15	0.0	138.0	0.69	742.0	18.0	2016-01-19 11:12:53.000	2016-01-19 11:15:10.000
4	98	5.0	178.0	0.25	902.0	15.0	2016-01-19 10:57:02.000	2016-01-19 10:59:59.000

endodometer	driverID	countryCode	startLocationCounty	endLocationCounty	countOfRawDataRecordsForTrip
NaN	243	NaN	estonia	estonia	120
NaN	243	NaN	estonia	estonia	143
NaN	300	NaN	sweden	sweden	0
NaN	300	NaN	sweden	sweden	0
NaN	300	NaN	sweden	sweden	7

Figure 6: Trips data

3.1.3 Crash data

Accidents information was given through the *Crash* table, which includes details about car(s) involved, fault and damages compensation; the table contained only 681 rows of different crashes.

	vehicleID	car	date	location	insurance	car1	fault1	comp1	car2	fault2	comp2
6	623	Vechicle VOLVO FH (623) has been involved in t...	2018-02-07 00:00:00.000	Iirimaa	If P&C Insurance AS	VOLVO FH, 2017	T�psustamisel	N�uet k�sitletakse	Unknown	undefined	undefined
7	57	Vechicle SCANIA R124 LA4X2NA 420 (57) has been...	2011-03-16 00:00:00.000	Other country	If P&C Insurance AS	VOLKSWAGEN POLO	0	Kokku Compensation over 2000 EUR	SCANIA R124 LA4X2NA 420, 2002	100%	undefined
8	618	Vechicle VOLVO FH (618) has been involved in i...	2018-06-04 00:00:00.000	Rootsi	If P&C Insurance AS	VOLVO FH, 2017	1	Vechicle damage was not compensated	SKODA FABIA	0%	Compensation over 2000 EUR
9	621	Vechicle VOLVO FH (621) has been involved in i...	2018-07-17 00:00:00.000	Tallinn, Harjumaa	AB "Lietuvos draudimas" Eesti filiaal	VOLVO FH, 2017	1	Vechicle damage was not compensated	MAZDA 6, 2006	0%	Compensation between 500 and 2000 EUR

Figure 7: Crash data

3.2 Observations in regard to available data

1. Most of the policy holders have no claim at all: in fact among the 681 rows about crashes there are also information about vehicle which have not been involved in accidents at all; only 234 vehicles are involved in 246 crashes, in a time interval from 1995 to 2019; thus only 34.36 % of Crash data could be effectively used to predict future crashes.

2. For some vehicles, the history information are limited which might create difficulties to determine future probability of crashes; this fact could be caused by recently insuring of the vehicle or driver, or recently added telematic devices.

3. Extreme high claim amounts are rare events, but they would have bigger impact to the total sum of damages, for each vehicle.

4. In the provided sample database, there is very limited information about claim damage amount, because it is provided through textual descriptions and ranges, thus no exact amounts.

5. There is a mismatch between trips and crash data, since not many trips rows are known for vehicles involved in accidents.

6. Some trips attributes seem in the first view not relevant to the main purpose (e.g., the field *stoppedAfter*, used to indicate the time interval the vehicle stays turned off).

3.3 Desired data

In order to improve both the quality of analysis and the meaningfulness of results, in this subsection we indicate which information would be advantageous to have stored in the database.

3.3.1 More about trips, crashes and cars

Trips rows are obtained aggregating several Raw data rows according to start and end locations, but a more proper aggregation method should be used to characterise better the trip types (e.g., distances, speeds, times for each kind of road respectively).

Trips data cannot be exploited appropriately if exact locations (i.e., GPS coordinates) of start and end points are not known; the field *countryCode* is not sufficient. The same happens for Crash data, where the *location* field is too much generic. Moreover, Crash data should be enriched with describing more detailed the moment of the accident, whereas up to now only information about calendar dates are indicated in rows. The accident time would be useful for instance to examine hotpoints in several period of the day.

More details about vehicles, such as car age and main use (e.g., passenger, taxi, truck, bus, van, etc.), should be accessible, together with the car (and driver) history (e.g., past accidents and issues).

3.3.2 Hotpoints

To examine better crashes and their relationships with peculiar streets or type of road, it would be helpful to retrieve information from *Open Street Map* (OSM) and other open source data. For instance, given a specific road:

- Datetime information of accidents (not only crashes of insured vehicles), with valuable granularity, considering period of the day, day of week, seasonality, etc.;
- Traffic information (e.g., number of trips/vehicles, number of braking, etc.);
- Population density (e.g., urban area, sub-urban area, countryside, etc.);
- Road information (e.g., surface, type, speed limit, cornerings, etc.).

4 Data Analysis and Visualization

This section, after listing some ideas we discussed before looking more deeply to data, illustrates some analysis we performed, to understand possible correlations between trips and crashes.

4.1 Brainstorming

The problem is interesting from different points of view. For end customers, exploiting information about their driving style could make them save money. Insurance policies would become more custom and insurers could know in advance potential drivers and cars more probable to be involved in future accidents. Last but not least, road safety is another key motivating reason to study this problem: in fact similar methods could be developed and used for fleet management and to promote training courses to drivers, in order to improve their driving.

First of all, there is the need of defining appropriate features to identify driving style, choosing among available and/or needed attributes from data.

The kind of information an insurance company would be willing to receive has to be established, in order to classify data input and output.

What could influence the outcome has to be understood adequately, for instance how repeating trips may affect the driving behavior, or going through some particular locations and hotpoints.

Possible connections between available attribute and crashes should be analyzed, computing existing correlations and functional dependencies.

Drivers could be partitioned into several classes, using clustering methods based on the driving style.

Simulation modelling also could be one of the tools to adopt.

Anyway, the first thing to do is to uniform data and build a meaningful dataset, starting from available information.

4.2 Preparation of the dataset

Aggregation and analysis have been performed using Microsoft Office Excel, Jupyter Notebook (Python) and WEKA, an open source software for data mining and machine learning.

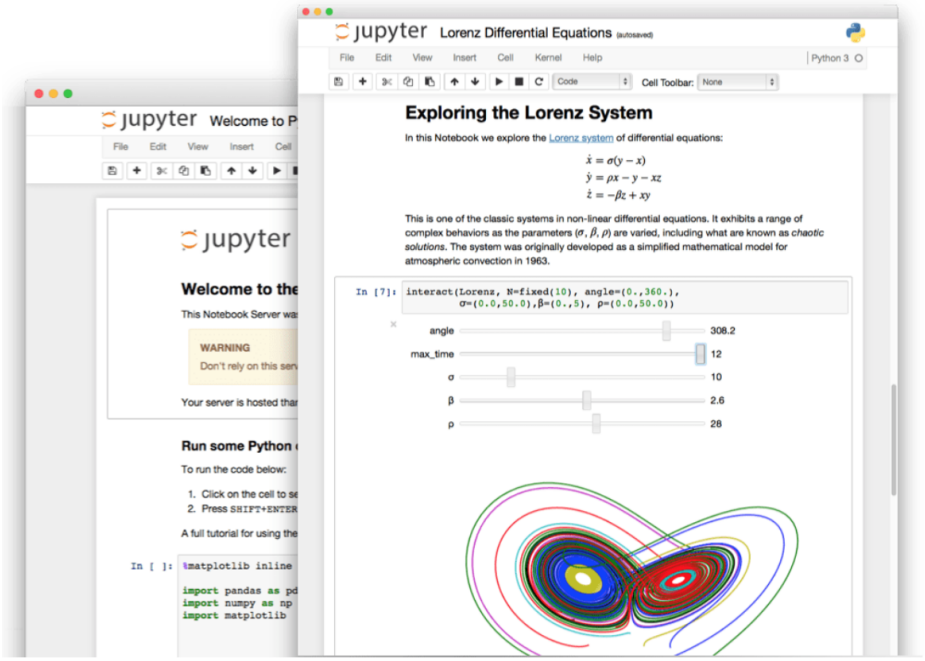

Figure 8: Jupyter Notebook

The database sample was located in Microsoft SQL Server.

Before putting together trips and crash data, we excluded some outliers values (e.g. in Trips data, rows with maximum speed of a general vehicle greater than 150 or average speed equal to 0 have been not considered).

Aggregation was done grouping rows by year, computing minimum, maximum and average values of distances, speeds and duration for each vehicle; about crashes, we summed up the number of crashed (with fault or not, respectively) in the same year for each vehicle.

The structure of the obtained data-set is illustrated in Fig. 9.

The last three columns, *number_faults_pred*, *number_nofaults_pred* and *avgDamages_pred*, use the crash information of the following year, if available, as dependent variable. These fields represent what we want to predict; in particular:

- *number_faults_pred* is the number of accidents where the fault is on the considered driver/vehicle;

vehicleID	year	distance_min	distance_max	distance_sum	distance_mean	maxSpeed_min	maxSpeed_max	maxSpeed_sum	...
1	2016	0.02	602.75	142583.49	69.349946	2.0	119.0	123748.0	...
1	2017	0.01	768.54	112783.69	65.955374	2.0	149.0	101286.0	...
1	2018	0.02	370.49	102716.07	54.433529	2.0	102.0	106151.0	...
2	2016	0.03	623.35	100801.50	64.041614	2.0	137.0	91273.0	...
2	2017	0.02	375.63	106713.66	60.222156	2.0	146.0	106934.0	...

avgMonthlyDistance	date	number_faults	number_nofaults	avgDamages	car_type	number_faults_pred	number_nofaults_pred	avgDamages_pred
11881.957500	0	0.0	0.0	0.0	0	0.0	0.0	0.0
9398.640833	0	0.0	0.0	0.0	0	0.0	0.0	0.0
8559.672500	0	0.0	0.0	0.0	0	NaN	NaN	NaN
8400.125000	2016-03-18 00:00:00.000	1.0	0.0	0.0	SCANIA R500	0.0	0.0	0.0
8892.805000	0	0.0	0.0	0.0	0	0.0	0.0	0.0

Figure 9: Aggregated trips and crash data

- *number_nofaults_pred* is the number of accidents where the driver was involved into but the fault was not his;
- *avgDamages_pred* indicates the yearly average amount for damages compensation where, because of lack of exact data, the damages were encoded by 0 (*no damage*), 1 (*damage below 500 EUR*), 2 (*damage between 500 and 2000 EUR*) and 3 (*damage over 2000 EUR*).

4.3 Analysis about available data

4.3.1 About trips

Plotting trips of all vehicles and examining only two features, average distance and average speed, it seems that drivers could be partitioned into two main groups, as it can be observed in Fig. 10:

- The first group including average distances at most equal to 40 km and average speed up to 50-60 km/h;
- The second group with distances between 50 and 150 km, with average speed between 20 and 50 km/h approximately.

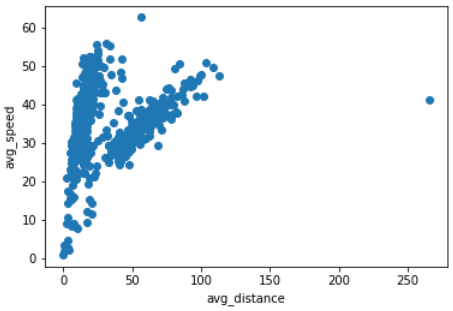

Figure 10: Scatter plot of trips and average distance and speed

From Fig. 10, it may be noticed that some outliers are still to be removed.

Two vehicles in detail

In Fig. 11 we chose to represent two particular vehicles, different from each other in their driving intensity, speaking about distance and average speed per day: in fact, Vehicle # 152 shows more homogeneous characteristics, whereas the behavior of Vehicle # 468 is more heterogeneous and unpredictable.

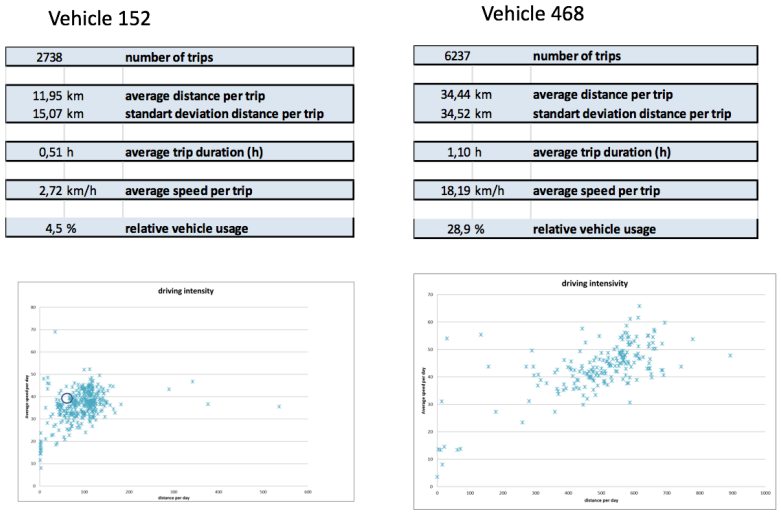

Figure 11: Details of two particular vehicles

4.3.2 About crashes

Grouping accidents by day of week, it can be noticed that 24.4% of crashes in the sample database happened on Mondays, as shown in Fig. 12.

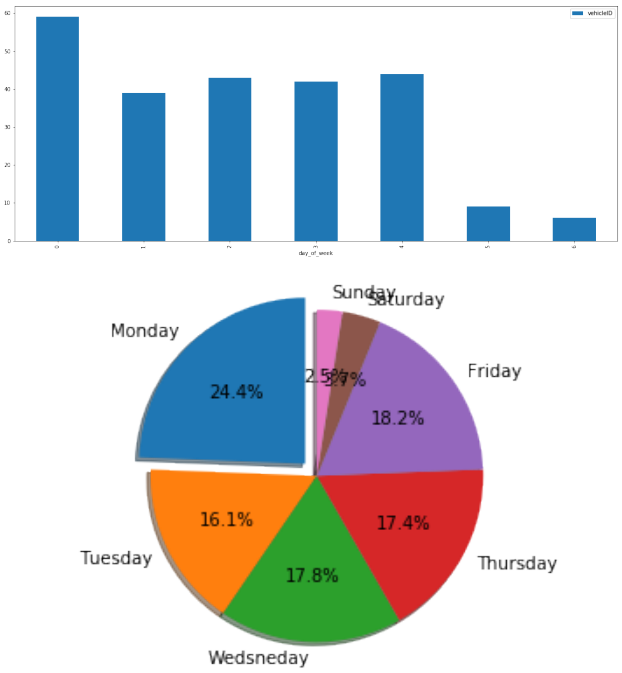

Figure 12: Crashes by day of week

Aggregating crashes instead by month of year, results show that there is no particular worse month when drive; anyway May presents a higher percentage of accidents, as seen in Fig. 13.

4.3.3 One day in details

Finally we show the driving intensity for Vehicle # 152 on a given date in Fig. 14.

During each day, every vehicle has its own patterns, in terms of distance and average speed, and the model to build needs to take it into consideration.

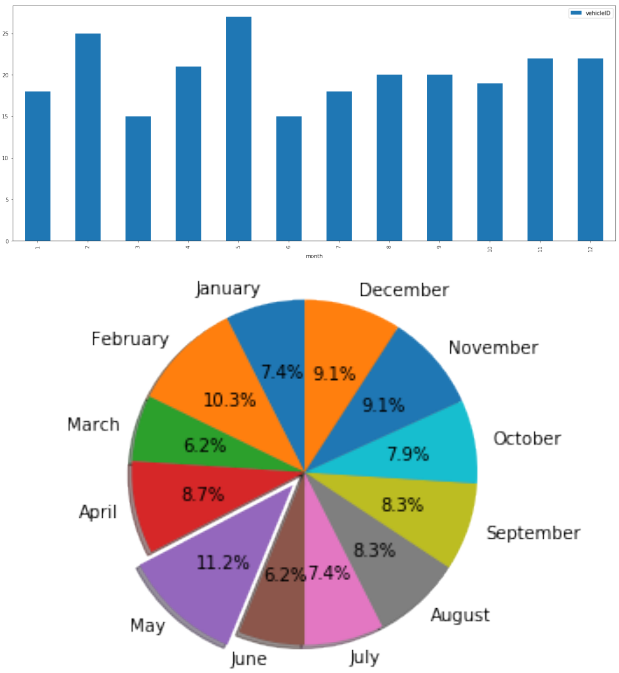

Figure 13: Crashes by month of year

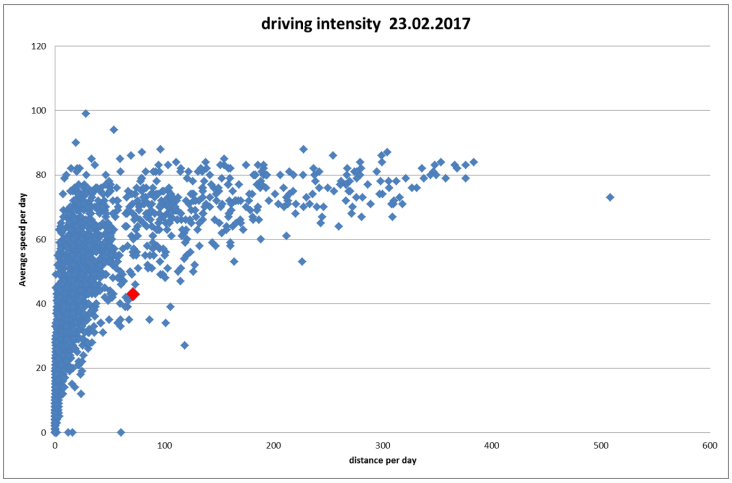

Figure 14: Distance and average speed for Vehicle # 152 on given date

5 Methodologies

This section describes the main approaches we applied to the given data; it also suggests a protocol to use to tackle the problem, explained through an example which could be extended and improved once desired data are collected.

5.1 Regression and Classification trees

Extremely Gradient Boosting (XGBoost), a Python scalable implementation of the Gradient Boosting, has been used for predictions and regressions on trips and crash data. It is based on an ensamble of *Classification And Regression Trees* (CART).

From the first tests, it was clear that this kind of model requires a more balanced dataset, i.e. more rows over the total number about crashes and trips of vehicles involved. Infact, according to the current data, when trying to predict the number of future crashes we obtained an average accuracy of 93%, using 10-fold cross validation in the sample dataset. However, the baseline prediction (i.e., obtaining no accidents as result) was 94%. This means that there are no significance differences among some vehicle characteristics that were or were not involved in accidents; thus, more features are needed.

Experiments were repeated for different attribute selection schemes (Feature ranking by correlation and information gain) and combinations of classifiers and regression algorithms (Logistic regression, Neural networks, XGBoost regression, K-means clustering etc.), with and without normalization and scaling of data, with other regression methods, but results obtained are similar: the *Root-Mean-Square-Error* (RMSE) or the accuracy of predicting no crashes is lower.

5.2 Locally Linear Embedding

Locally Linear Embedding is an unsupervised learning algorithm that computes low dimensional neighborhoods, preserving embedding of high dimensional data.

In fact, it is good for visualizing data points from multidimensional spaces into 2D plane, to give insight on the distribution of points in the hyperspace.

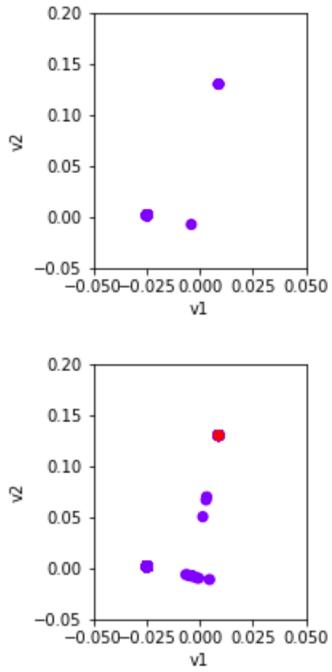

Figure 15: Locally Linear Embedding on sample dataset

We used it to confirm the findings from the classification. As it can be observed from Figure 15, there is visual overlap between the cases with accidents from the upper image and all cases without accidents in the lower image. This shows that the aggregated driving data from the Trips table give high overlap between risky and non-risky driving behavior. The intuition behind this is clear: Risky behavior does not always guarantee a crash but it only increases the probability of an accident happening. This is why an unsupervised, clustering method could yield a good approximation for defining risky behavior and finding which cars have higher probability of being involved in an accident, according to given data.

5.3 Suggested approach: Clustering

Given the structure of data and available and desired features, we suggest to adopt the following protocol to tackle the Usage-based insurance problem:

1. *Suggested approaches*:

- Using *Probabilistic Mixture Models* to model distributions of accidents and driver profiles;
- Using *Cluster centers* to identify typical behaviors within different groups.

2. *Compute*:
   - Occurrence of accidents in history;
   - Crashes distribution in the areas of interest and identify baseline insurance fee based on global risk of accidents.
3. *Identify individual behavior of participants*:
   - Use methods of clustering and probability distributions for obtaining the driving characteristics based on driving data;
   - Increase or decrease the baseline insurance fee, according to the calculated risk of the cluster the participant belongs to.

Note that the baseline insurance fee could be modeled as an *Index* with values in range [0,100].

5.3.1 Additional functions

Other things could be integrated, such as identifying driver's behavior and style not only when the policy has to be renews but during the whole policy period, in real-time or almost, in order to calculate his instantaneous risk.

If the driver's behavior moves from non-risk to risk group of drivers, notifications could be sent to both the driver and the insurer, in order to warn the former and to inform the latter.

5.3.2 Example

In this part we propose a simplified version of our approach, for now considering only two features (driven distance and average speed) in a given date, as seen in Fig. 16.

The blue dots represent trips of all vehicles happened in the given date (when an accident happened), whereas the red one indicates a crash which involved Vehicle # 505.

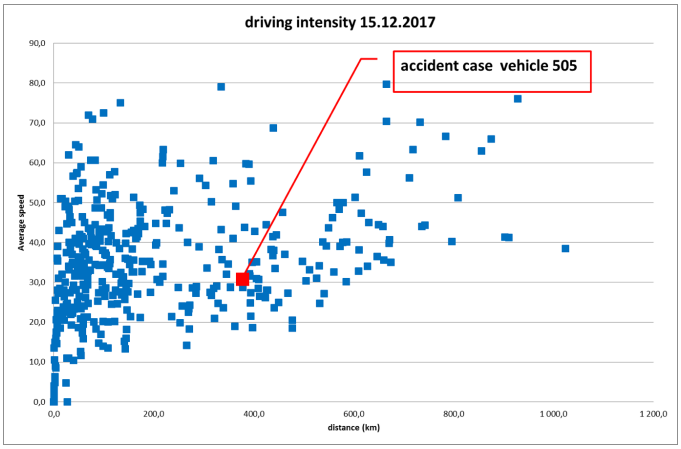

Figure 16: Suggested approach - Example with only two features

It is possible to compute the *distance* between the other vehicles and Vehicle # 505, considering distance in km and average speed in km/h during that particular day. Then it follows the computation of the probability to be involved into an accident, as shown in Fig. 17.

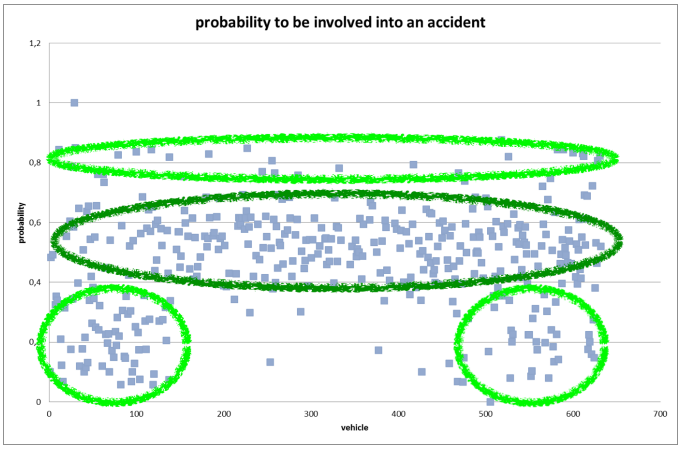

Figure 17: Suggested approach - Probability to be involved into an accident and Clusters

It can be noticed that four main groups are obtained, the one above with higher probability to be involved into an accident, always compared with Vehicle # 505.

Finally, starting from computed probabilities and clusters, it is possible to calculate the baseline PHYD insurance fee (i.e., the index with values in [0,100]) for each vehicle.

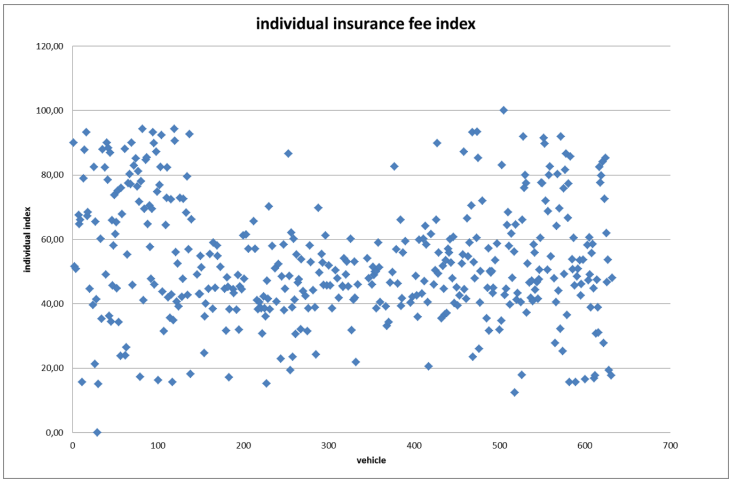

Figure 18: Suggested approach - Individual insurance fee index

6 Conclusions

In this final section we resume all the most important techniques and approach to use to tackle the Usage-based Insurance problem, answering to the three original questions asked by Holon Technologies; then we conclude proposing some ideas for further developments.

In this work we defined a possible method based on Probabilistic Mixture models and Clustering techniques to compute drivers' risk to be involved in future accidents.

6.1 Question 1

Which method and techniques could be used to classify the data?

To enrich the currently available information, we strongly recommend to retrieve Global Open Street Map data, together with the accidents history, to generate a baseline risk for each country or interested area.

Vehicles (and drivers) should be clustered based on their aggregated driving habits; to do this, we suggest to adopt unsupervised learning methods, which should suffice to the purpose of Question 1.

6.2 Question 2

How to find a (reasonable) trend for drivers' behavior and how to compute the probability of the trend to continue over time?

The approach mentioned in Subsection 5.3 can easily be updated as new data from the driver are available. Probabilistic Mixture models can be used to calculate the risk of the driver to occur into an accident, based on driving data; this risk index can be merged with the general risk of the areas the driver usually drives through.

Global models should also be updated based on the available data on certain periods of time (e.g., each month).

6.3 Question 3

How to define a model, combining different aspects, to determine the probability of damage, the damage frequency and the probable amount of the damage? Therefore how the premium should be changed? And what about pricing?

Based on the risk of accident calculated for answering to Question 2, we underline that *any function* can be used by insurance companies to calculate the premium as they prefer, since the output of our method is basically an index that insurers can exploit or customize to their own needs.

6.4 Further developments

The most urgent thing to do is to collect more data and specially more features, as indicated in Subsection 3.3. Once obtained this information, it would be possible to extend the simplified version described in Subsubsection 5.3.2, taking into account more than just two features.

Based on available data from fee calculations, a possible improvement of our method could be importing policies of different insurers to compute fees automatically and return, as output, the final value of premium directly. This would allow an overview of fees from both the customer side and the company side.

Another suitable way to try might be a graph-based approach. The model is based on spatio-temporal patterns, in particular on *close pair* of events. Similar approach (about predicting of burglaries on street networks) was presented in the work of Davies and Bishop [2]. Let $G = (V, E)$ be a graph where $V = \{v_1, v_2, \ldots, v_n\}$ is a nonempty set of n vertices (nodes) and $E \subseteq V \times V$ is a set of links (edges) which join them. The structure of this model can be represented with two graphs:

- One undirected graph, based on the concept of *spatial proximity* (i.e., the distance between two events);
- The other one, directed, built considering *temporal proximity* (i.e., the time period between two events).

Two events, indicated by two vertices in the graph, are connected if they occur in a given *threshold distance* or *time period*. More detailed, the events are vertices and the corresponding sets of links are defined in a different way, as follows, considering two threshold values, the *spatial radius* D and the *temporal radius* T:

$$E_d^D = \{(i,j) \mid d_{ij} \leq D\}$$

$$E_t^T = \{(i,j) \mid 0 < t_{ij} \leq T\}$$

If two events occur in a given time period T, they are connected by a directed link from the earlier event to the later. For that reason G_d^D is an undirected graph but G_t^T is a directed graph.

By analyzing both graphs, couples of events i and j which are close in space and time can be found. If two given events are close in both space and time then the corresponding vertices are adjacent. The *event graph* G_{dt}^{DT} of pairs which are close in space and time can be created (this new graph is directed). Practically, it is the intersection of the two graphs G_d^D and G_t^T. The set of links of G_{dt}^{DT} is:

$$E_{dt}^{DT} = \{(i,j) \mid (i,j) \in E_d^D \text{ and } (i,j) \in E_t^T\}$$

A structure like this could be exploited for finding hotpoints of accidents, given exact locations and datetimes of crashes. These hotpoints can be used by insurance companies in order to calculate the premium.

References

[1] Andreas Christmann. An approach to model complex high-dimensional insurance data. *Allgemeines Statistisches Archiv, December 2004, Volume 88, Issue 4, pp 375– 396*, 2004.

[2] Toby P. Davies and Steven R. Bishop. Modelling patterns of burglary on street networks. *Crime Science 2 (10)*, 2013.

[3] ESGI 151. First estonian study group with industry. `https://sisu.ut.ee/esgi151`, (last accessed on 2019-02-15).

[4] Guangyuan Gao, Shengwang Meng, and Mario V. Wuthrich. Claims frequency modeling using telematics car driving data. *Scandinavian Actuarial Journal*, 2018.

[5] Guangyuan Gao and Mario W. Wuthrich. Feature extraction from telematics car driving heatmaps. *European Actuarial Journal 8/2, 383-406*, 2018.

[6] Guangyuan Gao and Mario W. Wuthrich. Convolutional neural network classification of telematics car driving data. *Risks 2019, 7, 6*, 2019.

[7] Holon Technologies. `https://www.holontech.eu`, (last accessed on 2019-02-15).

[8] Marcos Marin-Galiano and Andreas Christmann. Insurance: an R-program to model insurance data. *Technical Report 2004, 49*, 2004.

[9] Mario W. Wuthrich. Covariate selection from telematics car driving data. *European Actuarial Journal, July 2017, Volume 7, Issue 1, pp 89–108*, 2017.