Anomaly Detection: Under the [data] hood in Smart Cars

Faisal Quader
Department of Information Systems
University of Maryland, Baltimore County
Baltimore, USA
fquader1@umbc.edu

Vandana P. Janeja
Department of Information Systems
University of Maryland, Baltimore County
Baltimore, USA
vjaneja@umbc.edu

Abstract - **This research focuses on discovering baseline models for driving behavior and vehicle functioning from data in smart cars. This facilitates detection of anomalous behaviors that deviate from such baselines. Human behavioral patterns capture frequent or repeated behaviors of users in the data. Here the data is from smart car sensors which capture driving behaviors as a direct function of how the vehicle responds to the use by drivers. We define models that represent these patterns in vehicle data to associate with the human behaviors for the detection of the anomalous situations in driving.**

We deal with different scales and resolutions of time where driver behavior is captured. We validate our findings of discovering behavioral baselines with frequent patterns discovered in the data. These computational models for driver behavior can provide baselines as well as help to discern truly adverse incidents. We can also apply these models to identify emerging cyber threats on smart cars.

Keywords – Anomaly Detection, Human Behavior, Driver Behavior, Clustering, baseline generation

1. INTRODUCTION

The safety of a vehicle depends on the mechanical state of the vehicle as well as the operator's driving pattern for a given time period. Driver behavior can make the driving patterns unusual due to speeding, sudden acceleration, irregular breaking, etc. along with mechanical failures. Therefore, human behavioral aspects can provide interesting insights into detecting anomalous activities in different settings to understand driver behavior and also distinguish driver behavior from mechanical failures. Such a distinction can also help inform future studies in determining baseline behaviors of drivers so that truly anomalous events originating in the car due to a mechanical or software failure based on vulnerability exploits can be discerned. In this work we mainly focus on driver behavior analysis and anomaly detection in driver behaviors.

In order to capture such behaviors, we analyze real time data from smart cars and detect any anomalous state to alert the drivers and the riders in real time. These heterogeneous but related data streams in a smart car are analyzed to come up with a wholistic approach to detect anomalies in the driver behavior. We present an alerting mechanism in smart cars to identify driver behavioral patterns. Specifically, we deal with different scales and resolutions of time where computational models for human behavior appear in heterogeneous manner. These computational models for driver behavior help us identify driver and vehicle performance.

1.1 Motivation

Vehicles (car, bus, fleet) are absolute necessity in our lives to go to places and/or move goods from one place to the other. We are also making our vehicles smart, fully connected to view real time traffic and weather, talk on the phone with Bluetooth, listen to the radio, watch video as well as getting real time status of automobile's mechanical functions. However, this smart interface comes with a price, which is the vulnerability to threats as well as malfunctions in mechanical parts of the vehicle. On the other hand the driver behavior can also be a point of vulnerability as far as safety is concerned. Driver behavior can be anomalous in response to mechanical failure or driver's own recklessness in driving. Identifying such patterns can also help create baseline behaviors so that truly unusual anomalies can be flagged.

We exploit the power of the connectivity within the vehicle along with machine learning to analyze the data streams generated in a smart car. We study driver behavior along with car telematics data in real time to detect anomalous behavior and alert drivers from getting into dangerous driving condition. We create a baseline for drivers and identify deviations from the baseline.

The rest of the paper is organized as follows. We discuss related work in section 2. Our methodology is outlined in section 3. We discuss our results in section 4 and finally in section 5 we discuss conclusions and future work.

2. RELATED WORK

Anomaly detection for smart cars is a gaining traction these days with more and more smart cars on the street and self-driving cars getting more and more attention. Foster et al. [1] talks about collecting telematics data from OBD2 (On-board Diagnostic) device that reads mechanical condition of the car and alerts telematic failure as it happens. Miller and Valasek showed how to simulate a cyber threat to the automobile remotely and take over the control of mechanical engine with break and acceleration [6]. Koscher et al. [5] demonstrate that an attacker is able to infiltrate virtually any Electronic Control Unit (ECU) and can leverage the ability to completely circumvent a broad array of safety-critical systems. This helps the car manufacturers taking care of those vulnerabilities for future models. Research shows that an automobile using a GPS based system can be vulnerable to different kinds of attacks, including blocking, jamming, spoofing, and physical attacks [3]. New car manufacturers are taking these findings and making sure the new automobiles are less susceptible to all these cyber threats. In such a climate of vulnerabilities, identifying driver behavior can help discern behavior from exploits of vulnerabilities.

Our research analyzes the temporal data for driver behavior and detects anomalous state of the driver as well as the anomaly from automobile's basic mechanical functions. Our model can also help detecting cyber threats on vehicles. We alert the driver in near real time, so the driver can take immediate action to perhaps mitigate the adverse effects to some extent. We bring in human behavioral aspects of the driver and detect the anomalous behavior during a defined time period.

3. METHODOLOGY

We study driver behavior on a defined time period and identify behavioral patterns. Consequently, time and human behavior are the common themes for anomaly detection in this research. We baseline driving pattern and any deviation from that to trigger alerts to flag the situation in near real time.

We utilized the approach as shown on the figure 1 below: First we collect car and driver data, then preprocess the data. We apply feature selection [2] to extract and focus on the key attributes since the data is very high dimensional and heterogeneous. Next, we do temporal binning to segment the datasets by time and apply clustering & classification based association rule mining to get to the final result of behavioral patterns and anomaly detection.

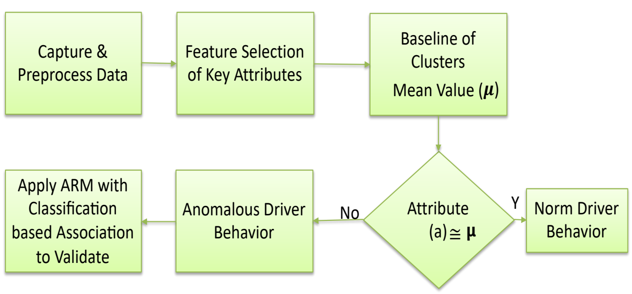

Figure 1: Driver Behavior Anomaly Detection

3.1 Data Collection

Driver behavior analysis along with capturing the mechanical state of the engine is done through a data collection mechanism from our automobiles. We collect data using an OBDII (On Board Diagnostic) based mobile application called Torque that collects telematics data in real time. OBDII device is connected to the automobile and the telematics data is collected from the device to the Android based mobile app via Bluetooth connection. Even though we have used Torque OBDII application for our research, there are many OBDII applications as possible alternatives to capture telematics data.

Figure 2 shows OBDII device and the Torque applications we used for our telematics data capture mechanism.

Figure 2: Telematics data Collection

We utilize our related work on developing an android based custom application (shown in figure 3) that captures raw OBD data attributes from the vehicle in real time and computes "derived" data like Sharp Turns, Sudden Acceleration, etc. as indicators of key driver behavior.

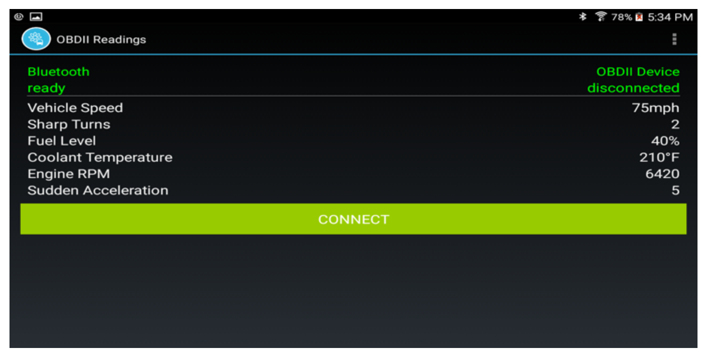

Figure 3: Custom Mobile App to Capture "derived" Driver data

3.2 Feature Selection

We captured numerous temporal attributes from the vehicle to study driver behavior. For example, just the OBD data generates close to 1500 attributes. In addition to the Telematics data through OBDII device, we also had the derived data like sharp turns, abrupt acceleration, abrupt deceleration, Engine RPM, etc. We applied feature selection to shortlist the key number of attributes. The objective of variable selection is three-fold:

- Improving the prediction performance
- Providing faster and more cost-effective predictors
- Providing a better understanding of the underlying process that generated the data. [2]

Therefore, in our research for driver behavior, feature extraction helps us narrowing down the key attributes among several data points captured from the vehicle and helps us identifying the anomalies in an efficient way.

Automatic feature selection methods can be used to build multiple models with different subsets of a dataset and identify attributes that can contribute to building an accurate model. We utilize the Recursive Feature Elimination or RFE with Random Forest algorithm on each iteration to evaluate our models. The algorithm is configured to explore all possible subsets of the attributes. The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is used to measure the differences between values (sample and population values) predicted by the model or an estimator and the values actually observed. We apply RMSE (Root Mean Square Error) as the Y-axis and the Variables in the x-axis. All 49 telematics attributes are selected here, and we observe that 9 attributes identified in figure 4 give almost comparable results individually. Therefore we can pick either one of these 9 attributes and eliminate the rest.

(2) "Longitude"
(11) "G.calibrated."
(19)"Distance.to.empty..Estimated..miles"
(24) "Fuel.flow.rate.minute.gal.min."
(29) "GPS.Latitude..."
(31) "GPS.Satellites"
(32) "Horsepower..At.the.wheels..hp."
(38) "Percentage.of.Highway.driving..."
(40) "Speed..GPS..mph."

Figure 4: Feature selection

remove attributes with an absolute correlation of 0.75 or higher. For instance, the variables which are highly correlated are:

(17) “Average.trip.speed.whilst.moving.only..mph.”

(38) “Percentage.of.Highway.driving... “

They have a correlation of 0.98349208. So one of the variables can be removed.

Figure 5 shows a matrix giving the correlations between all pairs of datasets. Interpretation of the Diagram and Ellipses are as follows:

The narrow Ellipses have high correlation.

The positive slope and blue are for positive correlations.

The bar in the right side represent the correlation values according to the color.

Using these rules, we eliminate extraneous attributes and focus on analyzing the key datasets for best accuracy and reliability.

Likewise, we can eliminate similar attributes as an attribute reduction strategy. For instance, we analyze the attributes from the OBD2 data and find the acceleration related attributes as shown in figure 6.

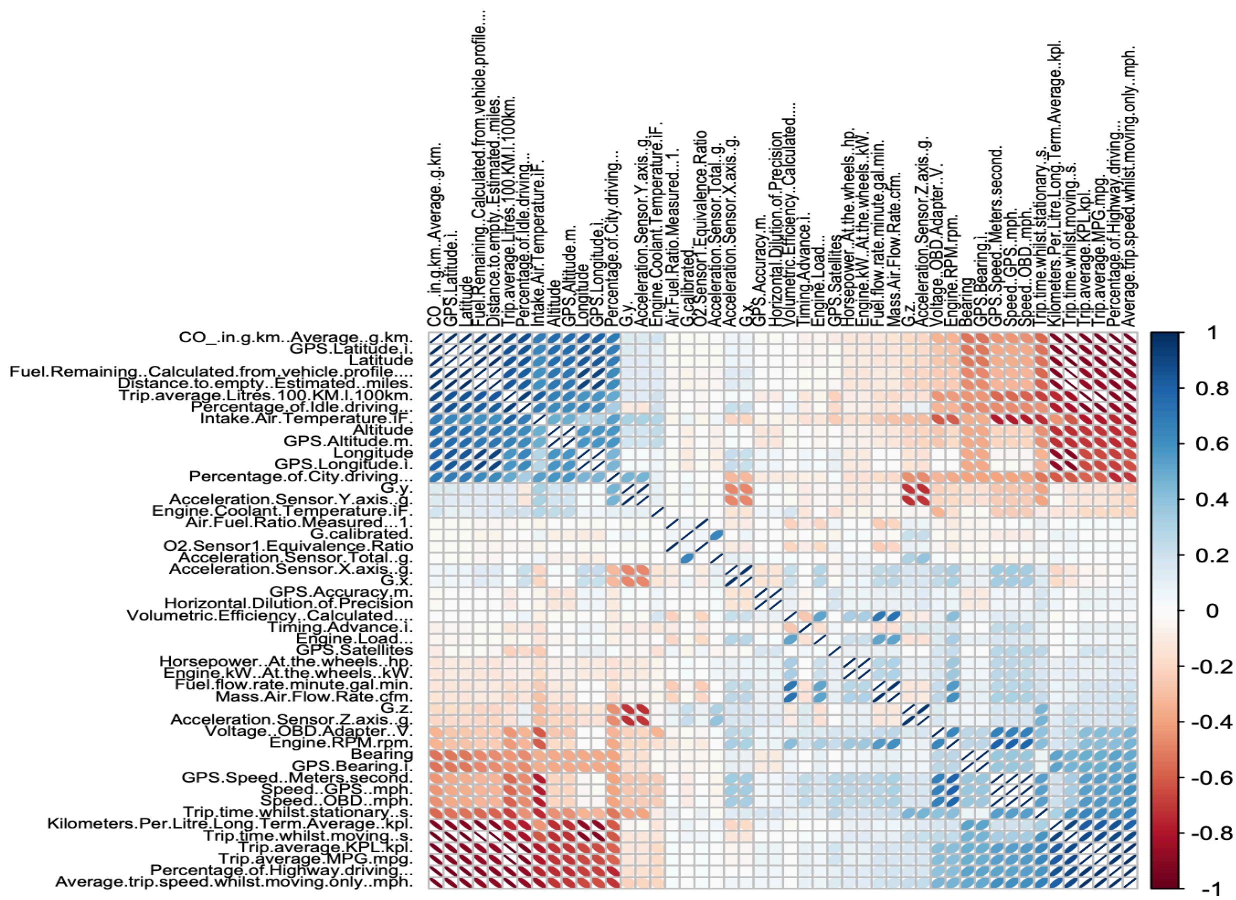

Figure 4: Correlation Matrix for All 49 Attributes

3.2.1 Correlation Matrix

Data may contain attributes that are highly correlated with each other. Many methods perform better if highly correlated attributes are removed. Generally, we want to

Acceleration related attributes :

(12)"Acceleration.Sensor.Total..g.",

(13) "Acceleration.Sensor.X.axis..g.",

(14)"Acceleration.Sensor.Y.axis..g.",
(15)"Acceleration.Sensor.Z.axis..g."

Related attributes:

(4)"GPS.Speed..Meters.second." and

(40)"Speed..GPS..mph." are similar to (16)Speed (OBD).

Lastly, (33)Trip speed is related to

(43)"Trip.average.KPL.kpl.",
(44)"Trip.average.Litres.100.KM.l.100km.",
(45)"Trip.average.MPG.mpg.",

(46)"Trip.time.whilst.moving..s."

(47)"Trip.time.whilst.stationary..s."

Figure 6: Acceleration attributes

Some of these are fairly recognizable and others are mostly related, so we can also perform simple ground truth validations on the feature sets as well.

3.2.2 Telematics and derived data attributes analysis by cars types:

We compare telematics and derived data attributes from 4 smart cars of mixed classes (Mercedes, BMW, Toyota and Honda) and highlighted the data attributes with common denominator. Age range of the 4 drivers is between 35 to 55 years, and we studied with 3 male and 1 female drivers. This choice of male and female was based on availability of the subjects and we plan to do more expansive tests including multiple participants in each category. Based on our analysis and the data coverage from all 4 cars, we take an intersection of attributes and focused on the intersection set of key attributes shown in table 1:

Table 1: Focused Key Attributes

Acceleration Sensor
Actual Engine % Torque – Engine RPM
Air Fuel Ratio
Avg Trip Speed
Engine Coolant Temperature
Engine RPM
Fuel Flow Rate/min
Fuel Remaining
Fuel Used
Speed (OBD)
Driver Behavior (Sharp Turn/Acceleration/Brake)
Engine speed / threshold violation – Engine Coolant Temperature
Low Battery – Voltage Adapter
Live weather and traffic updates
Real time Heart Rate Monitor - *Future*

3.3 Baseline of Driver Data

We create a baseline of the driver behavior and the vehicle's normal condition. We studied the 4 drivers for 3 months in different weather conditions like sunny, rainy, cloudy and snowy. We tried to do the experiments during the same period of time which is between 12 pm to 4 pm mostly in the city areas and a minimal highways. We use clustering to derive a baseline behavior. We use the elbow method to find the number of clusters as shown in figure 7. After identifying the K value, we perform K means clustering on the data and come up with a set of baseline clustering for each subject. We categorize the baseline clusters into the following 4 different types of safety condition ranging from safest all the way up to unsafe with two moderate conditions in between.

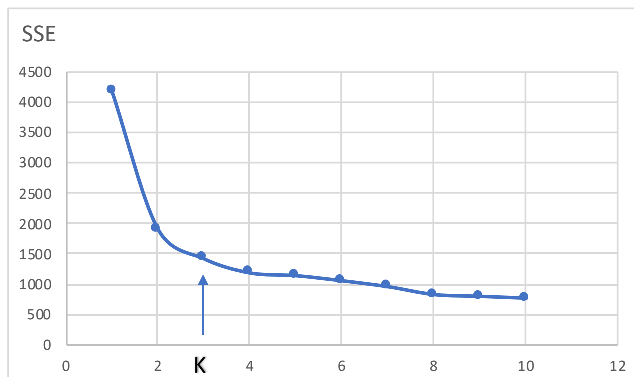

Figure 7: Elbow to determine the optimal number of clusters

3.4 Anomaly Detection of the Driver Behavior

We compare new driver data with their baseline clusters and categorize to the appropriate condition. If we find clusters outside of these conditions, then we trigger them as anomalous situations. We also applied Association Rule Mining on our vehicle data. We first converted the numeric data to categorical data through discretization. These ARM rules along with the Classification Based Association (CBA) rules help us drive rules that imply specific driving behavior based on parameters in the vehicle.

4. EXPERIMENTAL RESULTS

We will discuss the nature of the experiments, the datasets, the results and the Analysis & Validation in this section.

4.1 Datasets

First, we talk about the data manipulation of driver behavior. We have collected data on 4 different vehicles (2016 Toyota Corolla, 2005 Honda Odyssey, 2011 Mercedes Benz ML 350 SUV, 2014 BMW sedan i4) across different levels of luxury and non-luxury cars with 4 different drivers. We present our results on one of the more common vehicles (Toyota Corolla) and its baseline for one of the subjects (Middle aged). The subject data was collected under the conditions of day to day driving and no adverse tests were conducted.

4.2 Results

We collect data for a few days on regular time periods (between 4 PM & 6 PM) and append data in one dataset. Then we preprocess the dataset using feature selection methods. We cluster the selected features from telematics and driver behavior data. We picked k=4 as the ideal number from the elbow outcome as shown in figure 7. We then cluster the dataset into 4 clusters and identify the baselines from these clusters. We categorize the baseline according to the safety measures of the telematics attributes and driver behavior attributes. We finally compare the subject with this baseline for further comparison and analysis.

Following are the 4 clusters from the vehicle data:

Table 2: Clusters from Vehicle Data (Time of the day: between 4 PM & 6 PM)

Attributes	Full Data (3217)	Cluster 0 (716)	Cluster 1 (536)	Cluster 2 (994)	Cluster 3 (971)	Comment
Safety Categorization	*Overall*	*UNSAFE*	*SAFEST*	*SAFE*	*SAFER*	
Acceleration Sensor(Total)(g)	-0.0014	0	-0.0027	-0.0038	0.0006	
Air Fuel Ratio(Measured)(:1)	14.9527	14.7175	15.0472	14.8252	15.2044	
CO_ in g/km (Average)(g/km)	262.1087	319.812	294.1554	245.0225	219.3601	High emmision in cluster 0
Average trip speed(whilst moving only)(mph)	37.7857	13.194	31.7344	46.804	50.0275	Trip speed is high in cluser 3
Distance to empty (Estimated)(miles)	98.3274	110.6275	105.3646	96.6266	87.1139	
Engine Coolant Temperature(F)	193.5647	195.0961	192.4284	193.674	192.9508	
Engine Load(%)	34.0744	27.1301	38.7013	41.3587	29.184	High engine load in cluster 2
Engine RPM(rpm)	1840.6632	1109.2685	2146.6852	2315.1491	1725.3303	
Fuel flow rate/minute(gal/min)	0.0202	0.0115	0.0231	0.0288	0.0161	
Fuel Remaining (Calculated from vehicle profile)(%)	21.7179	24.5172	23.3882	21.3146	19.1447	
Fuel used (trip)(gal)	0.4984	0.0504	0.2312	0.5629	0.9103	
Intake Air Temperature(F)	101.0379	118.2668	96.7664	95.4111	96.4515	High temperature in cluster 0
Kilometers Per Litre(Long Term Average)(kpl)	14.4527	14.3978	14.3766	14.4662	14.5214	
Percentage of City driving(%)	43.0016	53.2712	51.0463	36.6352	37.5053	
Percentage of Highway driving(%)	32.9031	0	17.9611	46.6256	51.3661	
Speed (OBD)(mph)	43.6867	11.6151	57.4801	59.889	43.1357	High speed in cluster 2
Trip average MPG(mpg)	31.3372	25.7464	29.3618	33.4055	34.4329	
Voltage (OBD Adapter)(V)	12.6947	12.5162	12.7886	12.7728	12.6943	
Volumetric Efficiency (Calculated)(%)	43.0454	38.0922	44.8078	49.5563	39.0597	Low efficiency in cluster 0
Sharp Turns	0.0103	0.0377	0.0056	0	0.0031	High frequesncy in cluster 0
Sudden Acceleration	0.0096	0.0321	0	0.003	0.0051	High frequency in cluster 0

We also identified Association Rules in this dataset to derive the correlations between the attributes and identify the key rules to discover driver activities.

4.2.1 Baselines through clustering

We baseline the vehicle data and categorize the clusters by Safe, Safer, Safest and Unsafe types.

Cluster 0: This cluster is categorized as "UNSAFE" because of high CO emission, high air temperature, low volumetric efficiency, high frequency of sharp turns and high frequency of sudden accelerations. All these parameters are outside of the normal threshold and hence this cluster in labelled as UNSAFE.

Cluster 1: This cluster is categorized as "SAFEST" because of high efficiency driving patterns and very low to no incident of adverse driver behavior as well as the engine condition.

Cluster 2: This cluster is categorized as "SAFER" because of the majority of the attributes being in the safe range except occasional speeding and high engine load because of the speed.

Cluster 3: This cluster is categorized as "SAFE" because of most of the attributes being in the efficient range with the exception of occasional speeding as well as sharp turns and abrupt acceleration.

We compare this baseline with our subject (driver and the similar car – Toyota) for another instance of driving behavior for the same subject. We notice the anomalous behavior as shown in table 4.

Table 4: New Instance of Driver & Engine Behavior

Attributes	Full Data (1938)	Cluster 0 (1088)	Cluster 1 (147)	Cluster 2 (414)	Cluster 3 (289)	Comment
Safety Categorization	*Overall*	*UNSAFE*	*UNSAFE*	*UNSAFE*	*UNSAFE*	*ANOMALY*
CO_ in g/km (Average)(g/km)	243.5672	260.3344	70.0431	253.3363	254.7126	High emmision in cluster 0
Engine Coolant Temperature(F)	211.3373	213.5893	182.8844	210.2362	218.9094	Very high temperature in every cluster
Sharp Turns	0.2389	0.2243	0.3946	0.2802	0.1557	High frequesncy in every cluster
Sudden Acceleration	0.1615	0	0.1633	0	1	High frequency in cluster 3

Regarding this subject (driver and the automobile), we notice unusually high CO emission, unusually high engine coolant temperature, sharp turns as well as sudden acceleration. This is anomalous to the same driver's baseline clusters. We also run association rule mining as a validation of the findings through the clustering analysis.

4.2.2 Sample converted dataset:

We take the raw data from vehicles and discretize them with three different buckets with appropriate ranges. We generated histograms to identify the ranges for categorical data. For instance, engine coolant temperature fluctuates between 185 to 200 degrees. Therefore we divide the data attributes in 3 equal discrete intervals as shown in table 5.

Table 5: Numeric to Categorical Data Mapping

Engine Coolant Temperature(F) (Numeric)	Engine Coolant Temperature(F) (Categorical)
185	ECT 185 to 189.9
190.4	ECT 190 to 194.9
195.8	ECT 195 to 200

We have discretized the remaining data attributes the same way.

4.2.3 Analysis & Validation

We utilize association rule mining on the categorical data to identify the best rules with its associated attributes.

The rules shown in table 6 are anomalous (high measures) from driver behavior and also fall under the "Unsafe" cluster as discovered from the baseline cluster. Likewise the safe driving rules show low number of intensity which correctly fall under the "Safe" cluster of the baseline.

For instance, high air fuel ratio with high trip speed along with a high engine coolant temperature imply sudden acceleration. Likewise, high air fuel ratio with a very high CO emission imply sharp turns.

Table 6: Few Key Rules as Sample

Rules	*Confidence*	*Result*
Acceleration Sensor(Total)(g)=AS -0.07 to 0.2 ==> Sharp Turns=NO Sudden Acceleration =NO	0.97	**Low** Acceleration Sensor implies **no** sharp turn and **no** sudden acceleration
Average trip speed(mph)=ATS 34 to 51 Percentage of City driving(%)=PCD 33 to 65.9 ==> Sudden Acceleration =NO 1940	1.00	**Low** average trip speed with a **low** city driving implies **no** abrupt acceleration
Air Fuel Ratio=AFR 16.1 to 19.0 Average trip speed (mph)=ATS 34 to 51 Engine Coolant Temperature (F)=ECT 195 to 200.0 Percentage of Highway driving(%)=PHD 0 to 32.9 Speed (OBD)(mph)=Speed 50 to 75 ==> Sudden Acceleration =YES	0.93	**High** Air Fuel with relatively **high** speed with **high** Engine Coolant temperature would definitely imply Sudden Acceleration (**Yes**)
Average trip speed(mph)=ATS 34 to 51 Engine Coolant Temperature(F)=ECT 190 to 194.9 Percentage of Highway driving(%)=PHD 0 to 32.9 203 ==> Sudden Acceleration =YES	0.93	**High** trip speed with moderately **high** Engine temperature would imply sudden acceleration (**Yes**)
Air Fuel Ratio=AFR 16.1 to 19.0 CO_ in g/km (Average)(g/km)=CO 245 to 285.9	0.92	**High** Air Fuel with a very **high** CO

Thus the association rules also validate our findings through the baseline detection in clustering.

4.2.4 Real world Application Implementation

We utilized the findings from the methods developed in our research to help inform the development of a real world application. This work has helped with development of a mobile application called "School Bus Connect" for School buses and children who ride the bus. K-12 bus riders' safety depends on bus drivers' driving behavior and the condition of the bus. We monitor driver's behavior in real time and any deviation from the associated speed limit, sharp turns or abrupt acceleration, we raise alerts to the bus driver on the tablet application as well as send message to the school administrator's work station. This mobile platform also connects to the OBD2 device via Bluetooth and monitors vehicles mechanical state like the tire pressure, engine temperature, RPM etc. in real time so the system can raise alerts for any anomalous state on the vehicle.

4.2.5 Ethical considerations

While this paper does not directly address ethical questions as part of the methodology, future work can address and investigate the use of the massive amounts of data from smart cars to derive behavioral trends that can impact users adversely. For example, the patterns derived can impact insurance ratings of drivers, additionally if subjects are identified by demographics it may lead to propagation of biases that we see in society towards different types of drivers. Therefore, for such applications extensive experiments are required to validate the findings across multiple types of cars and multiple types of drivers with a clear goal of non-biased pattern discovery and to ensure fairness in the knowledge discovery process for this application. These are difficult questions to answer and formulate as studies. The authors want to recognize these as a key challenges for any studies in this area.

5. Conclusions and future work

In our research on driver behavior, we have identified a model to detect anomalous condition of the vehicle and also detect driver's behavior in near real time in comparison to baselines generated from historical data. We have also identified association rules between the key attributes of the vehicle and drivers data to validate our findings.

This research provides a grounding for future work in detecting anomalies resulting from the possibility of harms due to cyber threats, engine malfunction and/or driver's reckless behavior. The timeliness of the alerts in real-time is an important aspect to consider as well for future work.

Our research on driver behavior focuses on temporal range of a day. However, we have other key attributes on different time ranges that may directly impact the driver behavior on the road, such as mood of the driver, health condition of the driver, weather and road condition, etc. This research gives us a model to identify anomalies on human behavior based on few basic attributes. Additional models will need to be derived to add other impactful attributes as mentioned above.

6. REFERENCES

[1] Foster Ian, Prudhomme Andrew, Koscher Karl and Savage Stefan. Fast and Vulnerable: A Story of Telematic Failures. Department of Computer Science and Engineering University of California, San Diego.

[2] Guyon, Isabelle and Elisseeff Andr´e. An Introduction to Feature Extraction. March 2003. Springer Link.

[3] Hubaux Jean-Pierre, Apkun C Srdjan and Luo Jun. The Security and Privacy of Smart Vehicles. May-June 2004. In IEEE Security & Privacy.

[4] Icosystem. Cyber Security: Human Behavior Matters. Icosystem Corporation. 10 Fawcett St., Cambridge, MA 02138. Voice: (617) 520 1000. Available at http://www.icosystem.com/cyber-security-human-behavior-matters/.

[5] Koscher K, Czeskis A, Roesner F, Patel S, Kohno T, Checkoway S, McCoy D, Kantor B, Anderson D, Shacham H, and Savage S. "Experimental security analysis of a modern automobile". Evans D and Vigna G, editors, IEEE Symposium on Security and Privacy. IEEE Computer Society, May 2010.

[6] Miller Charlie, Valasek Chris. "Remote Exploitation of an Unaltered Passenger Vehicle". August 2015. Black Hat USA.

[7] 17 Monticello, M. & Shenhar, G (March 2016). CR Consumer Report. "Must-Have Car Features, and Those You Can Skip. Advice to help with optioning your next car". Retrieved from https://www.consumerreports.org/cro/news/2015/03/must-have-car-features-and-those-you-can-skip/index.htm