Consumer Experience With ADAS Towards Improving Road Safety

by Guangyu Xing, Jiwei Zeng, Peihsuan Hsia, Weihao Zeng, Xiancaozhi Yi

Masters of Science in Business Analytics, January 2019,
George Washington University School of Business

A Thesis submitted to

The Faculty of
The School of Business
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science
in Business Analytics

January 31, 2019

Thesis directed by:

Dr. Shivraj Kanungo

Chair of the Department of Decision Sciences
Faculty Director of Decision Sciences
Associate Professor of Decision Sciences & of Info Systems & Tech Management

Acknowledgements

We would like to thank Dr. Neeraj Koul and Dr. Ali Obaidi of the MITRE for providing the topic and for mentoring our team throughout this project. We would also like to thank Patrick Hall of the George Washington University School of Business for guiding our team on the data mining techniques used in this project.

Abstract

Consumer Experience With ADAS Towards Improving Road Safety

Our project is aimed to explore how to improve the function of Advanced Driver Assistant Systems (ADAS) based on the data of consumer experience from the National Highway Traffic Safety Association (NHTSA). And it is consisted of two parts with respect to: 1. building and developing the model to detect the records related to ADAS; 2. performing text mining in the filtered data and extract information. The team found that ……..

Executive Summary

Table of Contents

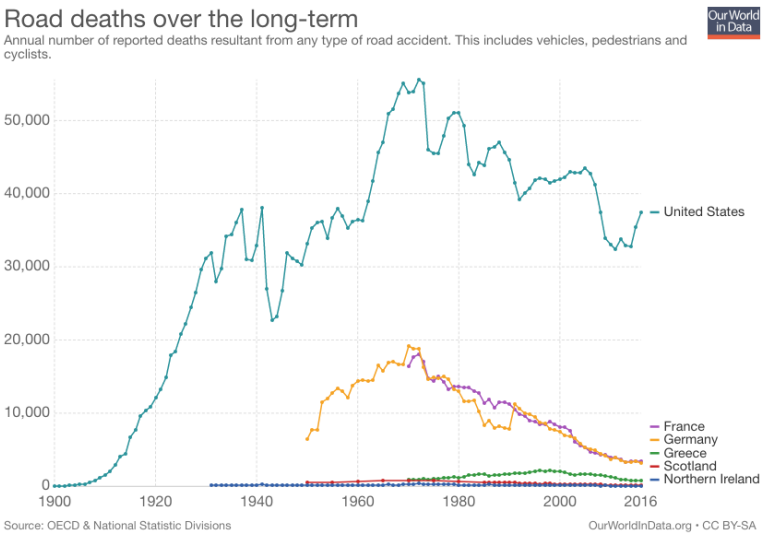
List of Figures

Chapter 1: Introduction and Background

Section 1.0: Introduction

Everyone who works for autonomous vehicles believes that human driving is not safe. They believe that autonomous vehicles can save human lives by eliminating thousands of preventable highway deaths each year. However, PHOENIX — A self-driving vehicle operated by Waymo was involved in a crash this June. According to existing media reports, the cause of the accident was that the safety driver equipped with the car fell asleep in the car. We cannot help thinking that are technologies really helping in improving road safety? And how can we improve these functions to eliminate the friction between human and machines' functions?

In this decade, many technologies are being widely used in the automotive industry. Advanced driver-assistance systems(ADAS) are the most trending technologies and have great market potential. When designed with a safe human-machine interface, ADAS should increase car safety and more generally road safety.



The concept of this project was provided by the MITRE Corporation, a federally funded research and development corporation, to produce an implementable data analytics solution about consumer experience with Advanced Driver Assistance Systems(ADAS) towards the goal of improving road safety.

The MITRE Corporation and National Highway Safety Association (NHTSA) are committed to improving the safety of drivers and vulnerable road users on our roads. Towards this purpose the project collects a relevant dataset about the safety-related complaints from NHTSA about interacting with ADAS system functions that can be used to inform analysis to improve safety. It is of interest to investigate how people interact with these functions.

Our project is aimed to find out any possible issues related to the ADAS and provide relevant suggestion on improving these systems. To fulfill this goal, firstly, we need to develop a classifier to essentially figure out the subset of report that are related to ADAS. After we get the subset of the data, we also want to figure out some trends in these complaints.

Section 1.1: Background

There is no doubt that vehicles have been one of the most important traffic styles in the United States. The United States is home to the second largest passenger vehicle market of any country in the world and the number of cars sold in the U.S. per year stood at 6.3 million in 2016. Also, there were an estimated 268.8 million registered vehicles in the United States in 2016, most of which were passenger vehicles. However, in 2016, 37,461 people died in motor vehicle crashes.

Advanced Driver Assistance Systems are aimed to assist drivers with the driving process. Research shows that the vast number of vehicle crashes are tied to human error and ADAS are designed to automate the process and minimize the human errors, to not only keep drivers and passengers safe, but they keep other drivers and pedestrians safe, too.

With more and more drivers learn about the concept of ADAS and automakers are constantly developing and implementing these new technologies, an increasing number of modern vehicles have ADAS. At the same time, the amount of complaints about ADAS is also increasing. Because the interaction between drivers and ADAS is one of the most significant factors, it is necessary to learn about these complaints and figure out solutions to improve the ADAS.

Section 1.2: Purpose and Scope

There are two purposes of this project. First, we build the classification model to classify whether or not the customer complaint related to ADAS. The classification model can tremendously increase the efficiency of finding the ADAS-related complaint.

By using this classification model, researchers do not need to manually screen each row of the complaint in order to find the complaint related to ADAS.

Second, after using the classification model to find the ADAS-related complaint. we use text mining method to further exam the details of complaint in order to see how drivers interact with ADAS system and what the specific problems related to ADAS system during the process of using it are. By text mining in the ADAS-related complaint, this text mining model can help the researchers know deeply about the ADAS system problems and further think of the solutions to solve the problem and enhance the quality of ADAS, making the car driving more safety.

Section 1.3: Exposition of Relevant Literature

There are a few previous studies focus on the database consumer complaint database from the NHTSA. The time trends show clear increases in complaints surrounding the Ford/Firestone tire recall and the Toyota unintended acceleration recall. Increases in complaints may be partially driven by these recall announcements and the associated media attention (Mahtab Ghazizadeh, Anthony D. McDonald & John D. Lee, 2014).

Advanced Driver Assistance Systems catch attention in academy. Their technical feasibility and possible obstacles were explored for road safety (Meng Lu, Kees Wevers & Rob Van Der Hejiden, 2005). Their functional potentials and limitations were also explored (J. Piao & M. McDonald, 2008). A survey of Pedestrian Detection was conducted, which led to fierce dicussion (David Geronimo, Antonio M. Lopez, Angel D. Sappa & Thorsten Graf, 2010)

Availability of Data

Chapter 2: Methods

Section 2.0: Availability of Data

In our project, we mainly focus on the database of consumer complaint database from the NHTSA which contains all safety-related defect complaints received by NHTSA since January 1, 1995. The database provides detailed information about the car, like the model, manufacturer and miles. Also, it includes the data about the accident or malfunction, such as the numbers of people injured or dead, the speed and location, and the original consumers' complaint.

Section 2.1: Classifying data -- labeled and unlabeled data

The original data does not discriminate whether the complaint belongs to ADAS and the subcategories of ADAS. Therefore, we first use random sampling method to sample 2500 complaint cases, manually labeled the complaint, and use the labeled data to construct models. We add 10 columns to the train data: "ADAS" (whether or not the complaint belongs to ADAS), "Adaptive Cruise Control" (whether or not the complaint belongs to one of the subcategories of ADAS - Adaptive Cruise Control), "Forward Collision Warning", "Lane Departure Warning", "Automatic Emergency Braking", "Blind Spot Monitoring", "Lane Keep System", "Rear Visibility Camera", "Adaptive Headlights", and "Park Assist". We then use "ADAS" as our target variable.

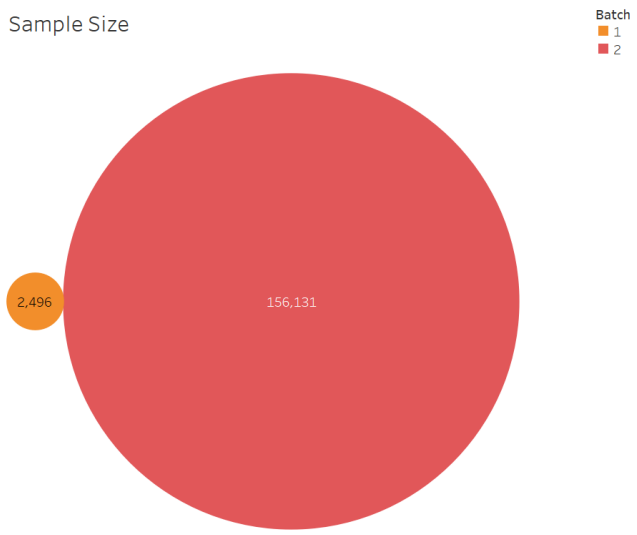
Section 2.2: Visualization of raw data

On data visualization, we visualized the 2500 labeled data and the rest of unlabeled data separately. By visualizing data, we not only can have the general concepts of data such as the number of deaths in the accident complaints but also can broadly

examine the representative of the 2500 sample data. The details of the visualization are as follow.

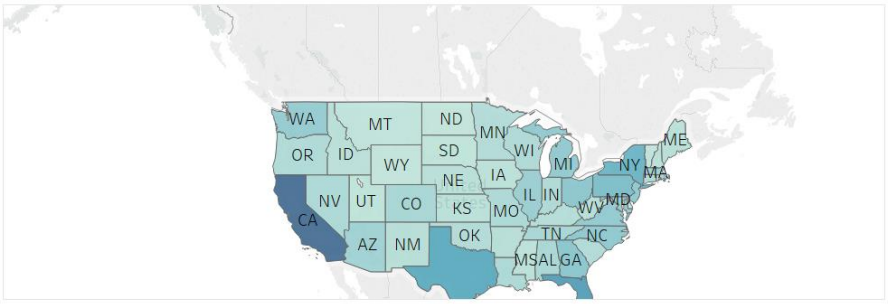
The labeled and unlabeled datasets are named as Batch 1 and Batch 2. Not only do we want to display the raw data by graph, but also, we are trying to make sure the patterns of Batch 1 and Batch 2 are similar. Therefore, it can be confirmed that we do the sampling randomly.

We can see that we have sample size of 2,496 in labeled data and 156,131 in unlabeled data.

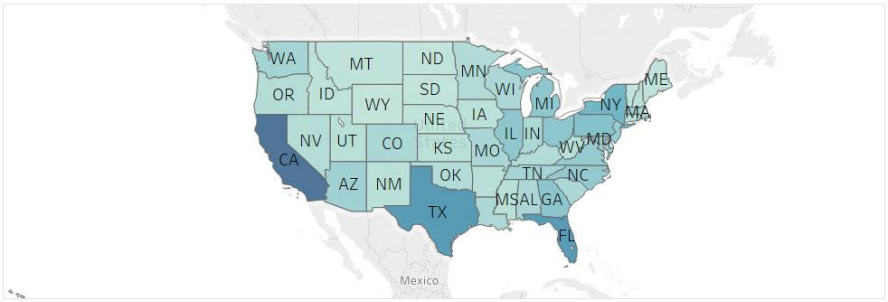


We analyzed geographical distribution of complaints and these two batches share the similar distribution. California, Texas, Florida and New York are the top 4 states of complaints. The complaints are mainly from the west and east coasts, which are positively relative to the population and the number of vehicles

Batch 1

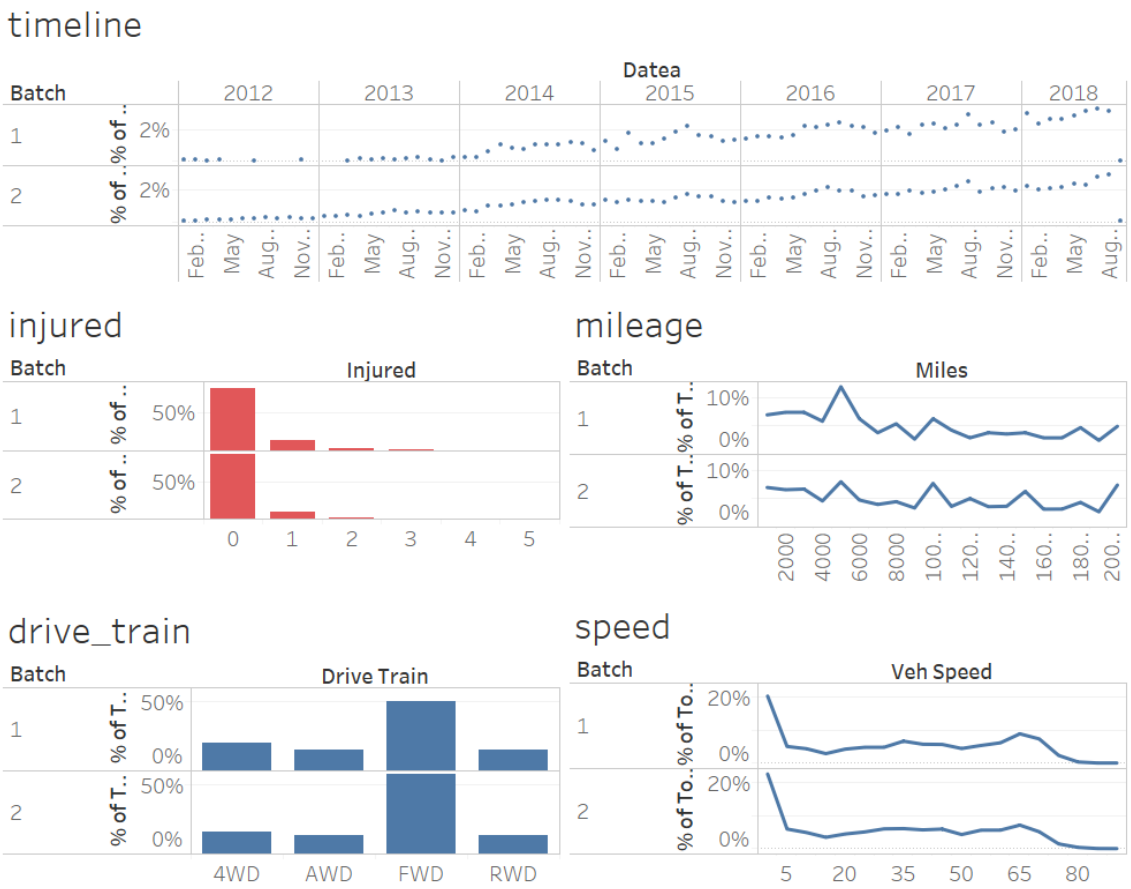


Batch 2



In the plots below, in the vertical axis is the proportion of the number of complaints.

From the time series of complaints, we can tell significant uprising trends in both of Batch 1 and 2. Also, it can be found that the peak of complaints occurred in each August and the though in winter. In almost each complaint, there is no injury. With the mileage increasing, there is no obvious uprising or downward trend for complaints. For drive train of vehicles, nearly 50% of vehicles are forward wheel drive cars. From the plot of vehicle speed and number of complaints, it is most common to have a complaint at speed of 65 mph.



For makes of vehicles, Ford, Chevrolet and Jeep contributed the most complaints. According to US Automotive Brand Rankings – YTD, Ford and Chevrolet are ranked at the first place and third place however Jeep is ranked at only sixth place, which means maybe vehicles of Jeep have more troubles.

In almost each complaint, there are no fire or death.

Section 2.3: Columns selection and generation

In this part, we will illustrate the columns we selected in the dataset and new columns we generated to create classifier models. We selected the columns that can have the relation to the ADAS systems and vehicle safety. The explanation of selection is below.

1.MFR_NAME (MANUFACTURER'S NAME):

Different manufacturers will have different quality control processes in the production, which generates diversified performance on safety in a long term.

2.MAKETXT (Make Name):

Each make/brand has its own branding positioning and target customers. Besides, each make manages its own dealer system, which will influence the maintenance in the after-sales.

3.MODELTXT (Model Name):

Vehicles in the same model share the same features and properties in safety. We concatenated the MAKETXT and original MODELTXT as the new MODELTXT because different makes can name the same model name.

4. YEARTXT:

Vehicle makes usually update models each year, which can improve the performance of vehicles or generate new problems. We concatenated the MAKETXT, MODELTXT and original YEARTXT columns to generate the new YEARTXT column.

5.crash/fire/injure/death:

These columns reflect the degree of severity of incidents.

6.COMPDESC (Component Description):

This column is the specific component involved in the incidents, which could highly related to some ADAS system.

7.CITY/STATE:

Difference road conditions and transportation regulations out of region can lead difference in driving habits, which influence the driving safety.

8.Miles/VEH_SPEED

Higher mileage and speed can cause more malfunction during driving.

9.cdescr

The most importance column.

10.Antibrake/Cruisecontrol

They can be highly related to ADAS, especially AEB and ACC.

We also generated several new columns to extract some information from raw data.

1.Age

We use FAILDATE and YEARTXT to estimate the age of vehicles.

2.Month/Weekday

We generate Month/Weekday from FAILDATE because the seasonality and weekday can influence the road condition and traffic.

Section 2.4: Data Cleaning

By using methods of handling outliers and missing data and of data encoding, we prepare the train data to construct models. First, we deal with the outliers of 4 important columns -- “INJURED” (number of persons injured in the car accident), “DEATHS” (number of persons dead in the car accident), “OCCURRENCES” (number of the malfunction occurring), “VEH_SPEED” (the speed of the car when the car accident happened). For “INJURED” and “DEATHS”, since a maximum accommodation for a car/minivan is 8 people, if the number of injured or dead people is over 10, we adjust it to 8. For “OCCURRENCES”, we choose 95th quantile as our boundary. If the number of occurrences is over the value of 95th quantile, we adjust it to the value of 95th quantile. As for “VEH_SPEED”, if the speed of the car exceeds 160 km/h, we adjust it to 160 km/h.

Second, we deal with missing values in 9 columns -- “INJURED”, “DEATHS”, “OCCURRENCES”, “MEDICAL_ATTN” (whether or not the medical attention was required), “VEHICLES_TOWED_YN” (whether or not the vehicle was towed was), “POLICE_RPT_YN” (whether or not the incident was reported to police), “ORIG_OWNER_YN” (whether or not the car was driven by its original owner), “ANTI_BRAKES_YN” (whether or not the anti-lock brakes was on when the incident happened), “CRUISE_CONT_YN” (whether or not the cruise control system was on when the incident happened). For “INJURED” and “DEATHS”, we use 0 to fill the null value since we do not know the exact number of injured/dead persons. For “OCCURRENCES”, we use 1 to fill the null value since we think that the malfunction should happen at least one time to cause the incident. For all of the binary variables -- “MEDICAL_ATTN”, “VEHICLES_TOWED_YN”, “VEHICLES_TOWED_YN”, “POLICE_RPT_YN”, “ORIG_OWNER_YN”, “ANTI_BRAKES_YN” and “CRUISE_CONT_YN” -- we use “N” to fill the null value since we are not sure the exact condition.

After handling missing value, we use dummy variables, also called one-hot encoding, to process the categorical variables.

Section 2.5: Analysis Approach

Predictive model building process:

After we clean the dataset, we are going to utilize sklearn to build predictive model.

Since we have a very limited data set, we want to make full use of the data by using 5 - fold cross-validation. To avoid data leakage in the cross - validation process, pipeline along with RandomizedSearchCV in sklearn can make sure in each cross - validation loop, all data preprocess steps are fitted on the training, and the data transformation are performed on all dataset. The following is the working flow of our model building process.

1. Initialize RandomizedSearchCV
2. Start an iteration and randomly select a set of pre - defined hyper parameters
3. In each cv loop:
    a. impute missing value
    b. preprocess non-text column and text column

        (i. and ii. are perform at the same time using feature union)

        i. extract all non-text columns, then use z-score to standardize numeric values for each column
        ii. extract the text column, then run TfidfVectorizer on this column
    c. build gradient boosting machine
4. Return to step 2 until all iterations are done
5. Select the best set of hyper parameters according to the grid_scores, then refit on the whole dataset to get the final model.

Glossary of Terms

References

Ghazizadeh, M., McDonald, A. D., & Lee, J. D. (2014) Text Mining to Decipher Free-Response Consumer Complaints. Human Factors: The Journal of the Human Factors and Ergonomics Society, 56(6), 1189-1203.

Lu, M., Wevers, K. & Van Der Hejiden, R. (2005) Technical Feasibility of Advanced Driver Assistance Systems (ADAS) for Road Traffic Safety. Transportation Planning and Technology, 28(3), 167-187.

Piao, J. & McDonald, M. (2008) Advanced Driver Assistance Systems from Autonomous to Cooperative Approach. Transport Reviews, 28(3), 659-684.

Geronimo, D., Lopez, A. M., Sappa, A. D. & Graf, T. (2009) Survey of Pedestrian Detection for Advanced Driver Assistance Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(7), 1239 – 1258.

The MITRE Corporation. (n.d.). Retrieved October 28, 2018, from https://www.mitre.org/

The National Highway Traffic Safety Association (NHTSA). (n.d.). Retrieved October 28, 2018, from https://www.nhtsa.gov/

Wikipedia. (n.d.). Retrieved October 28, 2018, from https://www.wikipedia.org/

(n.d.). Retrieved October 28, 2018, from https://www.safercar.gov/Vehicle-Shoppers/Safety-Technology/AEB/aeb

(n.d.). Retrieved October 28, 2018, from https://owner.ford.com/support/how-tos/safety/driver-assist-technology/driving/how-to-use-lane-keeping-system.html

(n.d.). Retrieved October 28, 2018, from http://www.goodcarbadcar.net/2018/08/u-s-auto-sales-brand-rankings-july-2018-ytd/

Appendix

The Mitre Corporation (MITRE)

The Mitre Corporation, a non-profit organization, is managing federally funded research and development centers (FFRDCs) supporting several U.S. government agencies, including Department of Defense, Department of Homeland Security, National Institute of Standards and Technology and so on. In addition, MITRE is responsible for maintenance of the Common Vulnerabilities and Exposures (CVE) system, which is the industry standard as a reference-method for publicly known information-security vulnerabilities and exposures, and the Common Weakness Enumeration (CWE) project.

Since MITRE's mission is to assist the government with the safety, stability, and well-being of our nation, it is interested in enhancing the road safety by improving the Advanced Driver Assistance Systems (ADAS). In reality, it is drivers who always interact with the ADAS, so performing research on the drivers' or customers' experience about cars can help us find out problems about ADAS and provide related suggestions on advancing ADAS.

National Highway Traffic Safety Administration (NHTSA)

The National Highway Traffic Safety Administration is an agency of the Department of Transportation with mission as "Save lives, prevent injuries, reduce vehicle-related crashes." As part of the Corporate Average Fuel Economy (CAFE) system, NHTSA is responsible for writing and developing and Federal Motor Vehicle Safety Standards regulations for motor vehicle theft resistance and fuel economy. Also, the NHTSA is undertaking the duty of the maintenance of the data files for traffic safety research.

Advanced Driver Assistance Systems (ADAS)

In our project, we will focus on nine kinds of ADAS, Adaptive Cruise Control, Forward Collision Warning, Lane Departure Warning, Automatic Emergency Braking, Blind Spot Monitoring, Lane Keep System, Rear Visibility Camera, Adaptive Headlights and Park Assist.

Adaptive Cruise Control, is an optional cruise control system for road vehicles that automatically adjusts the vehicle speed to maintain a safe distance from vehicles ahead. Control is based on sensor information from on-board sensors.

Forward Collision Warning, is an automobile safety system designed to prevent or reduce the severity of a collision. It uses radar and sometimes laser and camera to detect an imminent crash.

Lane Departure Warning, is a mechanism designed to warn the driver when the vehicle begins to move out of its lane (unless a turn signal is on in that direction) on freeways and arterial roads.

Automatic Emergency Braking, detects an impending forward crash with another vehicle in time to avoid or mitigate the crash. These systems first alert the driver to take corrective action to avoid the crash. If the driver's response is not sufficient to avoid the crash, the AEB system may automatically apply the brakes to assist in preventing or reducing the severity of a crash.

Blind Spot Monitoring, uses digital camera imaging technology or radar sensor technology to detect vehicles in adjacent lanes and warn drivers of approaching vehicles. These systems are most effective when drivers are passing other cars, being passed, or

making a lane change. Blind spot monitor systems are an option on many new cars, SUVs and trucks and can help you avoid a crash.

Lane Keep System, uses a camera mounted behind the windshield's rear view mirror to monitor road lane markings and detect unintentional drifting toward the outside of a lane. If the camera detects an impending unintentional drift, the system will use the steering system and the instrument cluster display to alert and/or aid you to stay in your lane.

Rear Visibility Camera, is a special type of video camera that is produced specifically for the purpose of being attached to the rear of a vehicle to aid in backing up, and to alleviate the rear blind spot. It is specifically designed to avoid a backup collision.

Adaptive Headlights, is the idea of moving or optimizing the headlight beam in response not only to vehicular steering and suspension dynamics, but also to ambient weather and visibility conditions, vehicle speed, and road curvature and contour.

Park Assist, is an autonomous car-maneuvering system that moves a vehicle from a traffic lane into a parking spot to perform parallel, perpendicular, or angle parking.