Prediction of Driving Outcomes Based on Driver Behaviour and Roadway Information

Ybralem Bugusa
Mtech Student, Department of Computer Science and Engineering
Symbiosis International University, Pune, India
ybralem.wekele@sitpune.edu.in

Prof. Shruti Patil
Assistant Professor, Department of Computer Science and Engineering
Symbiosis International University, Pune, India
shruti.patil@sitpune.edu.in

Abstract- Intelligent Transportation System (ITS) is advanced transportation framework for the more quick-witted utilization of transportation system. The setup consists of an arrangement of coordinates propelled advancements which utilize hardware, computers, correspondence, sensor, and cameras to upraise travelers with relevant information for the safety and competence of transportation system. Advanced Driver Assistance System (ADAS) is a piece of ITS, that helps drivers in the driving procedure to empower safe condition. In such a system, to guarantee the wellbeing of all the street clients, it is important to upgrade the execution of Advanced Driver Assistance System (ADAS) for lower classes of vehicles. Real-time driving safety risk prediction is a primary component of an ADAS. In this paper, we are going to use a dataset from Secondary Strategic Highway Research Program (SHRP2) Naturalistic Driving Study. The elastic net regularized multinomial logistic regression model is going to be utilized to predict the traffic safety risk that can be updated to the individual driver based on specific variables. The prediction performance of the model will be improved by using re-sampling techniques.

Keywords: traffic safety, driving outcome, naturalistic driving behavior, multinomial logistic regression, elastic net, re-sampling, advanced driver assistance system.

1. INTRODUCTION

Having the growing technologies in intelligent transportation, improving the performance of Advanced Driver Assistance System (ADAS) is necessary to ensure road safety. Modern ADAS are systems that help drivers in real time driving process by warning to the driver or actuation of the control systems to reduce traffic collisions that cause injury, death of pedestrians as well as drivers and property damage.

Developing a model that predicts the real-time road safety risk (driving outcome) for ADAS is important in creating a safer driving environment for drivers, passengers, and pedestrians. For all of this, it is important to have a real-time data of the traffic safety factors, i.e., vehicle and other infrastructure device information with regard time, location and other relevant information in association with the main safety factors of driver behavior, roadway information, and environmental condition data. The aim of this paper is developing model that predicts the real-time driving outcome of driver during driving situation that can be updated to individual driver information. It will use Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS) data to make a predictive model of real-time traffic risk for ADAS. The model uses some of the data above by aggregating the data of individual trip.

To classify the driving outcome of the driver using multiple traffic safety factor, the elastic net regularized multinomial logistic regression [1] will be used due to its ability work with correlated variable and bias/variance trade-off.

The remaining of this work is arranged as follows: First, we provide the related work with a brief description and evidence about SHRP 2 data and the influence of driver behavior in traffic safety risk prevention. Finally, we are going to discuss proposed work and conclusion.
II. LITERATURE REVIEW

a. Naturalistic Driving Study data

Traffic damages are an expanding group medical problem, too much touching touchy groups of street and poor people. Moreover, half of the general population executed in vehicle accidents are youthful grown-ups elderly in the age of 15 and 44 years – regularly the representatives in a family and the cutting edge the next generation which will receive their country. Furthermore, Traffic accident Cost developing countries 1% to 2% of their gross national commodity – more than the aggregate development of these countries. [21]

To overcome this distraction many analysis and crash risk level prediction model have been done by evaluating the crash casual factors data to provide vital information for effective transportation policy, vehicle design, driver awareness and potential countermeasures. However, there have been certain criticized in data observation studies of crash casual factors due to poor quality data; absence of available and valuable exposure measures linked to the observation; the incomparability of self-revealed safety-related events; and, the difficulty in evaluating accountability for the safety-related events [2].

[5] Paper uses traffic data from loop detector to develop prediction models of real-time crash risk for different segment types and states of traffic flow on Korean expressways. They address the result difference in occupancy and level of overcrowding play a significant role in the ramp vicinities and level of congestion difference between the downstream and upstream leads to a hazardous condition. However, it has the limitation of lack of weather-related data and lane-based traffic data.

[12] paper takes into account the traffic data from loop detector data and weather data from web crawler with the objective of investigating predictors of future traffic crash involvement. They handle bias of data for former drivers and compare characteristics of old drivers and crash-involved drivers. In the study, it includes only older adults living in Kashiwa. Therefore, the sample is biased toward active, healthy older adults who could take part in social activities. The sample size is small.

The second Strategic Highway Research Program (SHRP2) is research method studying the largest and most comprehensive Naturalistic Driving Study (NDS) ever undertaken linked with Roadway Information Database (RID). Initially, the 100-Car Naturalistic Driving Study collect the large-scale NDS data in the US which is first instrumented-vehicle study funded by the National Highway Traffic Safety Administration (NHTSA) and the Virginia Department of Transportation (VDOT). Then it is followed by a larger and more wide-ranging study, the Strategic Highway Research Program 2 (SHRP2) conducted from 2006 to 2015 holds data about driver behavior, road, vehicle, and weather and traffic conditions with age 16-68 in the event of either crash or near-crash.

The NDS and RID data allow researchers to associate driving behavior to specific roadway conditions for better analysis and traffic risk prediction. Detail report of SHRP2 NDS data instruments, data dictionary, all variables of the traffic safety factors and how to query and filter states in [11] [18-19] papers.
In this study, we are going to use the SHRP2 NDS particular variable of the safety factors, i.e., driver behavior, road, vehicle, and whether information due to several reasons [2]:

- Have advanced data collection techniques for different safety factor
- No information biases and
- Quality information primarily to identify the driver fault in near-crash/crash events.

b. Driver behavior

Recently, many kinds of research work in the prediction of real-time risk to prevent the sudden loss of human life by car and other vehicle and acknowledged that driver behavior plays a significant role in driving risk [1]. It presents that nearly 90% of the light-vehicle crashes included a similar sort of human error such as impaired conditions, accidental mistakes, and risky driving behavior and driver mistake is additionally a center explanation behind roughly 87% of every single business vehicle crashes. Since factors associated with individual driver is different [10], developing models that can be customizable to the individual driver is essential to making good countermeasures to traffic safety risk. Driving behavior can be more influential when it is associated with roadway, vehicle and weather condition data.

In [5] paper the authors develop real-time crash risk prediction models for different segment types and traffic flow states on Korean expressways using Genetic programming technique.

Authors [20] investigate the potential contribution of three predictors (i.e., driver anger, impulsiveness, and aggressiveness) of aggressive and transgressive manners on the road using multiple regression analysis models.

The [6] develop a model for detecting the driver’s involvement likelihood in secondary tasks from distinctive attributes of driving behavior (i.e., lateral acceleration, longitudinal acceleration, throttle position, and yaw rate). Using supervised feed-forward Artificial Neural Network (ANN) finds that the Selected driving performance attributes were useful in detecting the associated secondary tasks with driving behavior.

In [14] the author uses driver behavior and physiological features for crash prediction to advanced vehicle collision avoidance system using supervised learning model.

Several studies use different data set to explore the relationship between the involvement and the traffic safety-related events and considers input data used for their work. However, due to several reasons they aware some improvements on data sample usage, some of them they don’t use weather-related data and lane-based traffic data [5]. Others they may not evaluate variables that influence the performance of prediction [20] (i.e., years of driving experience is not considered).

In this study, we are going to use the driver behavior as a major contributor to safety risk with external factors that may influence it, i.e., roadway information weather condition, etc.
of SHRP2 NDS data which is the vast research method ever conducted. It contains data from vehicles, road, and weather condition data about driver behavior. We will consider some specific variables from all this data that are important for the predictor to classify the real-time driving the outcome of the individual driver.

Many studies have been conducted by using driver behavior with different factors to predict safety-related traffic problems. There are also studies which donot consider other factors rather than driver behavior.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Factors used to predict trafficsafety risk</th>
<th>Classification Algorithm</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>Driver behavior along with roadside information</td>
<td>Elastic net</td>
<td>68.5%</td>
</tr>
<tr>
<td>[2]</td>
<td>Vehicle ROR safety-related data</td>
<td>Bayesian Multivariate Poisson Log Normal Model</td>
<td>97%</td>
</tr>
<tr>
<td>[3]</td>
<td>Driver behavior, vehicular and road data are used</td>
<td>Classification and Regression Tree (CART)</td>
<td>66%</td>
</tr>
<tr>
<td>[6]</td>
<td>Vehicle-speed, vehicle longitudinal and lateral acceleration, yaw rate, and throttle position of the vehicle to predict secondary task of the driver.</td>
<td>Artificial Neural Network (ANN)</td>
<td>99.5%</td>
</tr>
<tr>
<td>[9]</td>
<td>Combination of driver glance behavior and traffic density</td>
<td>Linear Regression</td>
<td>-</td>
</tr>
<tr>
<td>[10]</td>
<td>It uses personality, demographic and driving behavior data</td>
<td>Negative Binomial (NB) Model and Logistic Regression</td>
<td>-</td>
</tr>
<tr>
<td>[12]</td>
<td>Driver behavior and crash history</td>
<td>A Multinomial Logistic Regression</td>
<td>-</td>
</tr>
<tr>
<td>[14]</td>
<td>Driver behaviors and vehicle information</td>
<td>Discriminant Analysis</td>
<td>95.25%</td>
</tr>
</tbody>
</table>

Paper [3], [6], [9] and [10]’s has considered the additional factors like vehicle information and other factors. But, they have used a limited number of sample data which can affect the accuracy of the system. In paper [14], the driver data was collected from simulator driving. In this case, the accuracy can be less when it is compared to the real driving situation.

Paper [1] uses weather and vehicle information, in addition to driver behavior and roadside information which can be more accurate in identifying safety-related traffic problems.

![Fig 2: Factors affecting traffic safety risk](image-url)
To improve the accuracy of [1], we will use resampling technique.

III. PROPOSED APPROACH

a. Data preparation and sampling

For further pre-processing of the SHRP2NDS data set, we will use Data munging and wrangle for cleaning and punctuation correction. It also helps to check the structure and value of the data according to the data dictionary; this contributes to map the data into R. R is the tool that we will use in this study. To handle the missing value, we are going to use Random Forest package to fill the missing value which is the fastest approach [26]. Multivariate Imputation by Chained Equations (MICE) is an expensive method used to handle the missing value [27].

We have collected the data from baseline video reduced data and Event video reduced data. Baseline video reduced dataset consists of 19,600 baseline events with 27 variables which includes detailed epochs, driver state, and driving environment information derived from video reduction. Event video reduced data consists of 68 crashes or 760 near-crashes with 57 variables which includes detailed event, driver state, and driving environment information derived from video reduction. By joining the dataset into Baseline, Near-crash and Crash events, we have selected total 21 important variables which includes driver state, and driving environment information. Hence the number of Baseline event data is extremely large. Therefore, we have applied random sampling techniques on major class (i.e. baseline event data) to reduce the data size required to achieve better classification. After data cleaning, 3353 total data are left which consists of 2595-Baseline (B), 969-NearCrash (N) and 62-Crash (C) Event Datasets.

![Graph showing the reduced dataset distribution](image)

Fig 3: Reduced dataset

Various applications can overcome the problem of imbalanced data between classes. The rightness of prediction can be improved when it handle the imbalanced data. The occurrence of imbalanced data is generalized to one-sided towards the larger part class. In any case, in a few applications, the accuracy of prediction in minority class is additionally noteworthy as much as in majority class. The accuracy in minority class will be increased by using re-sampling techniques. Such as under-sampling, over-sampling and bootstrapping are the most used techniques. Under-sampling and over-sampling are re-sampling based methods which don’t consider how the data are scattered in the space [25].

Bootstrapping is statistical resampling strategy to appraise the accuracy of the sample by utilizing subsets of accessible data or drawing haphazardly with substitution from a set of data points or original data.

In this study, we are going to use bootstrapping method to increase the accuracy of the classifier and overcome such problem.

b. Driving outcome classification

The driving outcome or risk events of SHRP2 NDS are into three crash, near-crash baseline or normal state. Near-crash is conflict situation requiring a rapid and careful maneuver to avoid an accident.

The author [1] further decomposes the events into 5 class or outcome to better create a safe environment for pedestrians as
well as driver based on the unsafe driver behavior and secondary task that the driver involved. Accordingly, we have decomposed the dataset into Baseline1, Baseline2, Baseline3, Near-crash and Crash.

TABLE 2: THE DECOMPOSED OUTCOME AND THEIR MEANING

<table>
<thead>
<tr>
<th>No.</th>
<th>Outcome</th>
<th>If it is</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Baseline1(normal driving1)</td>
<td>Minimum risk</td>
</tr>
<tr>
<td>2</td>
<td>Baseline2(normal driving2)</td>
<td>Self-hazardous</td>
</tr>
<tr>
<td>3</td>
<td>Baseline3(normal driving3)</td>
<td>Hazardous to self and others</td>
</tr>
<tr>
<td>4</td>
<td>Near crash</td>
<td>Near crash</td>
</tr>
<tr>
<td>5</td>
<td>Crash</td>
<td>Crash</td>
</tr>
</tbody>
</table>

The risk events were three which are showing by color spectrum for normal (baseline) state is “green,” for near crash “yellow” and “red” for crash event.

![Fig4: Driving outcome before classification](image)

To create safer environment which is especially useful for young drivers as well as drivers with limited driving ability. The baseline event is divided into three based on the driver involvement in unsafe driving behavior (i.e. improper turn, driving slowly in relation to other vehicle) and secondary task(i.e. texting, cellphone call, eye glance) that the driver involves during driving situation.
This can be done by selecting the prevailing unsafe driving behavior and secondary task from the dataset. The decomposed one is show in table2, two colors are added to represent the level of risk baseline1 is “green” as it is, baseline2 is “green-yellow,” and baseline3 is “khaki.” The remaining near-crash and crash event color is as it is. For more, we have shown in below figure 5.

The following figure illustrates the number of each driving outcomes after classification.

---

**Fig5: Driving outcome architecture**

**Fig 6: Driving outcome after classification**
In this paper, we propose weighted regularized multinomial logistic regression (MLR) which is precisely known elastic net [23] for variable selection and classification.

**Multinomial Logistic Regression** is the type of logistic regression technique used to evaluate the parameters of a multivariate explanatory model used when the dependent variable is binary (0/1, true/false and yes/no) and the independent variable is continuous or categorical [8].

It is the type of logistic regression with more than two outcomes but not ordinal. Hence we have more than two outcomes we will use the multinomial logistic regression. However, it’s hard to work with a highly correlated independent variable to overcome it will collaborate with a regularized method that is weighted regularized multinomial logistic regression. Elastic net is a merging of L1 and L2 regularization.

**LASSO regression** is the type of regression technique that works with highly correlated independent variable solves the high variable relation by shrinkage parameter lambda which takes one variable and shrinks the rest to zero that is the problem with multicollinearity. It is the L1 regularization [23-24].

In reverse, the **Ridge regression** is L2 regularization solves the multicollinearity by shrinking the value of coefficients but doesn’t reach to zero [23-24].

Elastic net is a group of Lasso and Ridge Logistic Regression. While we are working with high dimensional, correlated features, other reasons that improve the accuracy and performance of prediction are:

**Reasons why we will use elastic net:**
- Integrated variable selection,
- Capability to work with bias/variance trade-off,
- Cost-sensitive classification,
- Capacity to handle highly correlated variables,
- Ability to deal with high-dimensional problems, and
- Work with both numerical and categorical variables

Problems in prediction can happen due to either biased or variance; It may also occur by both. Regularized regression techniques have their own mechanism to overcome this:

---

**System Architecture**

---

Fig 7: Architecture of proposed work.
TABLE 3: FORMULATION OF ELASTIC NET

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Allow where n &lt; p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Solve multi-collinearity by minimizing coefficients but not to zero.</td>
</tr>
<tr>
<td></td>
<td>L2 norm based penalty.</td>
</tr>
</tbody>
</table>

**Formula**

$$\hat{\beta}^{ridge} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} u_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right\}$$

$$\hat{\beta}^{ridge} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} u_{ij} \beta_j)^2 \right\} \text{ subject to } \sum_{j=1}^{k} \beta_j^2 \leq t$$

**Lasso**

**Advantages**

- Allow where n < p
- Similar with ridge but multi-collinearity by shrinking coefficients to zero.
- L1 norm based penalty.
- Feature selection

**Disadvantages**

- Cannot perform group selection

**Formula**

$$\hat{\beta}^{lasso} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} u_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{k} \beta_j \right\}$$

$$\hat{\beta}^{lasso} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} u_{ij} \beta_j)^2 \right\} \text{ subject to } \sum_{j=1}^{k} \beta_j \leq t$$

**Elastic Net**

**Advantages**

- Allow where n < p
- It is shrinkage method.
- Uses both L1 and L2 norm based penalty.

**Formula**

$$\hat{\beta}^{elastic} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} u_{ij} \beta_j)^2 + \lambda_1 \sum_{j=1}^{k} \beta_j + \lambda_2 \sum_{j=1}^{k} \beta_j \right\}$$

$$\hat{\beta}^{elastic} = \min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} u_{ij} \beta_j)^2 \right\} + \lambda_1 \sum_{j=1}^{k} \beta_j + \lambda_2 \sum_{j=1}^{k} \beta_j$$

*Note: $\lambda_1 = 0, \lambda_2 > 0$ is equivalent to ridge regression and $\lambda_1 > 0, \lambda_2 = 0$ is equivalent to the LASSO. If both $\lambda_1 > 0, \lambda_2 > 0$ it is Elastic Net.*

**Table 4: Comparison of Elastic Net Classifier with Other.**

<table>
<thead>
<tr>
<th>Features/classification algorithm</th>
<th>SVM</th>
<th>Ridge Logistic Regression</th>
<th>Lasso Logistic Regression</th>
<th>Elastic net</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performing automatic Variable selection</td>
<td>NO</td>
<td>NO</td>
<td>Can’t complete group selection</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Handling of Bias/variance trade-off | NO | Yes | Yes | Yes
--- | --- | --- | --- | ---
Dealing with Correlated data and high dimension data | NO | NO | NO | YES

Elastic net is supervised classification model used in many areas for variable selection, optimization and classification purpose [22] [24].

**Elastic net prediction model for driving outcome**

Suppose: driver i, trip j, location l and time t.

**Procedure:**

1. $Y_{ijkl} \in \{1,2,3,4,5\}$;
   - 1, 2, 3, 4 and 5 indicates the driving outcome.
   - Such as Baseline 1, Baseline 2, Baseline 3, Near-crash and Crash respectively.
2. Consider the independent safety factors represented by $Z_{ijkl}$
   - $Z_{ijkl}$ includes the target driver, vehicle, and all other factors such as roadway information, historical crash events, weather information.
3. Hence, the function $f(Z_{ijkl}Y_{ijkl})$ shows the relationship between independent (predictor) and dependent (outcomes) variables
4. Divide the safety variables based on the regularity of real-time changes:
   - Time-series variables, $T_{ijkl(t)}$,
   - Event variables, $e_{ijkl(t)}$ Where
     - $T$ = constant time step,
     - $N$ = length of time-series
5. Redefine location and time of the predictor as $Z_{ijkl(t)}$.
6. Define the conditional probability of the driving outcome as the following:
   - $P_{outcome} = P(Y_{ijkl}=outcome \mid Z_{ijkl(t)}) = f(Z_{ijkl(t)})$.
7. The driving outcome to be predicted given $Z_{ijkl(t)}$ is computed as:
   - $Y_{ijkl} = \text{Arg max} \quad P_{outcome} \quad \text{Outcome}=1,2,3,4,5$.
8. Based on above formulation, we can apply elastic-net model to this work.
   - Number of class k = 5,
   - Outcome
     - 1 = baseline 1,
     - 2 = baseline 2,
     - 3 = baseline 3,
     - 4 = near-crash,
     - 5 = crash

$\hat{\beta}_{\text{elastic}} = \arg\min_{\beta} \left( ||Y - X\beta||_2^2 + \lambda_2||\beta||_2^2 + \lambda_1||\beta||_1 \right)$

- Naïve elastic net penalty
  - $P_{a} = \sum_{i=1}^{n} \left( 1 - a \right) h_{ij}^2 + a |h_{ij}|$
    - $f \text{ for elastic net, } 0 < a < 1$
  - But, it has the problem of double shrinkage, with the same a level
- To overcome this problem, elastic net uses,
  - Penalty = $(1 - a)\beta^2 + a|\beta|^2$
  - $= \lambda_2||\beta||_2^2 + \lambda_1|\beta|_1$ where $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$
  - $\hat{\beta}_{\text{elastic}} = (1 + \lambda_2)\hat{\beta} (\text{naive elastic net})$

**IV. CONCLUSION**

Prediction of road traffic risk during driving process plays an important role in saving of human life as well as infrastructure damage. Since, driving outcome of the driver can be influenced by internal and external factors, it is being better to develop predictive model that can be customizable for individual driver. In this investigation, we have offered elastic-net regularized multinomial logistic regression for prediction and classification of driving outcome and we are going to use the SHRP2 NDS dataset to achieve this. As improvements, we are going to use re-sampling technique to improve the performance of the model. For further extension of this work, using best search algorithm and advanced machine learning algorithm are possible to improve its performance.

**REFERENCES**


