List of Figures

DESCRIPTION OF THE PROBLEM

The problem of interest is to build a predictive model to estimate a composite traffic safety risk measure that changes temporally and spatially, and takes into account driver behavior, roadway quality conditions and historical safety characteristics of roadways. A subsequent problem formulation will use this predictive model to sample safety risks at some equidistant time epochs during a trip. These sampled risks are then compiled into a risk profile for that trip. We are also interested in the statistical analysis of these risk profiles to better understand one's driving patterns, and by devising appropriate filtering schemes, we like to correlate these driving patterns to location, time, weather and other useful parameters. These risk profiles, if aggregated over location parameters, can also give us useful information on the road safety from an infrastructure point of view. Last but not least, safety risk predictions can be used toward the development of Advanced Driver Assistance System (ADAS).

APPROACH

Prior to this project, our team had experimented with 100-car driver's behavioral study and had built a predictive model using the data that is publically available from this study. The funding from this project was used to acquire a larger dataset that is available from publically available SHARP2 database. We did not have sufficient funds to acquire the whole dataset, and for that reason we had to employ a filtering mechanism that closely matched our predication requirements. Despite our best efforts in devising the filtering scheme we were not able to obtain all the data that was necessary for our analysis. For instance, it was quite important for us to receive video data especially at times of accidents (crash or near miss). But, due to security and confidentiality reasons, this data was not made available to us. Our funds was not sufficient either to obtain a dataset that contained trips per a defined set of drivers, hence, some aspects of driver behavior could not be included in our modeling scheme. Our dataset, however, included baseline, near miss and crash data for random drivers, stamped with time and events. It included time series data in order of 10 -15 seconds prior to an event (e.g., crash) and 10 seconds after the event. This allowed us to extract real useful behavioral information at the time of events. The new dataset enabled us to significantly expand our safety risk prediction model and to develop a new model that explains driver reaction time on the basis using kinematics and statistical data. This is an important finding, which allows one to customize driver alerts according to driver's own characteristics and surrounding circumstances. Computation of reaction time and safety risk require a database architecture that was also planned and developed under this project.

METHODOLOGY

Our work pursued the following objectives:

- Development of appropriate data models and a database using multiple data streams. We used the data fields from Naturalistic Driver study and combined it with data from Straight Line Diagram (SLD) and historical crash data. Figure 1 illustrates the general data framework used for our data modeling. We adopted a feature engineering technique to prepare data from model building in the next step.

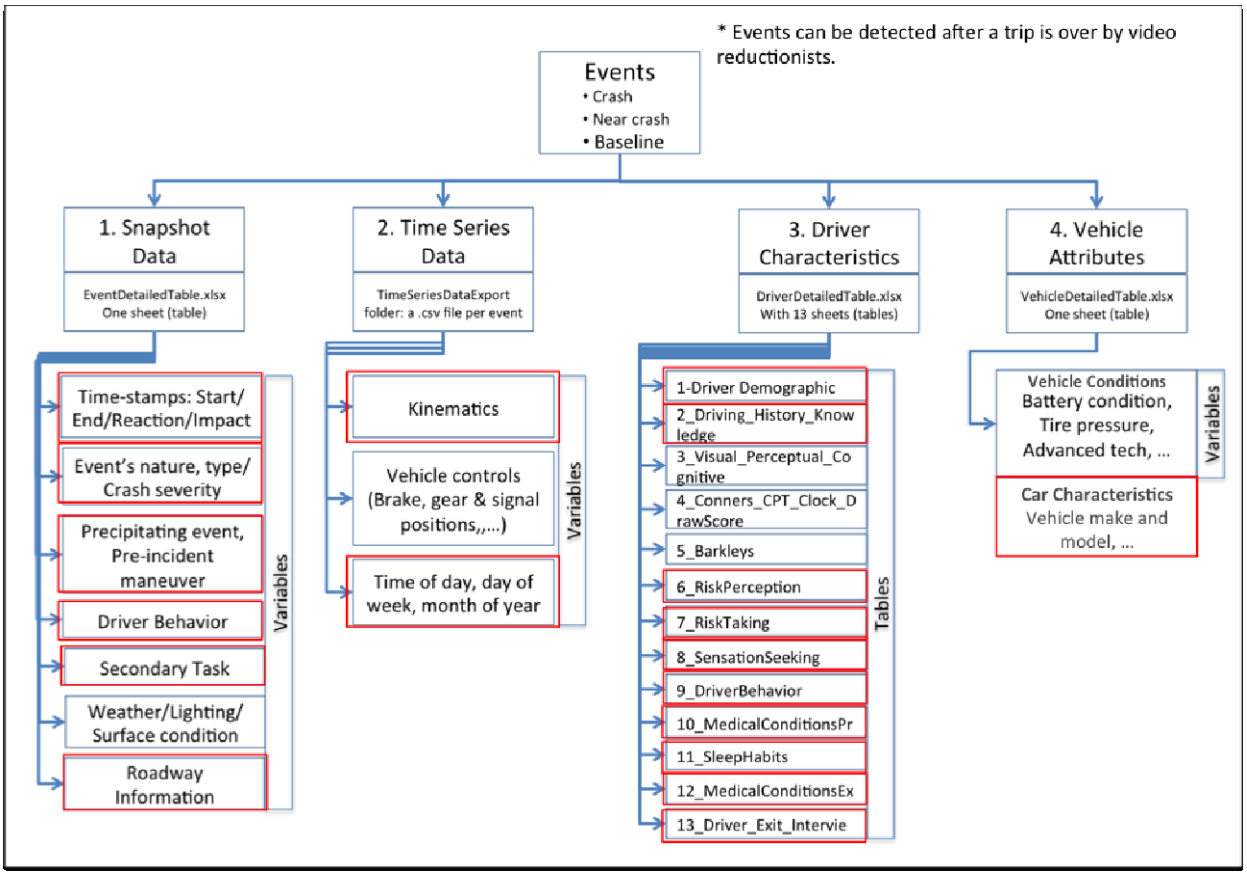

Figure 1 – A General Data Collection Model – "red boxes are primary candidates"

• Development of a safety risk predictive model – This model is intended to predict safety risk score of a driver at sampled time epochs. The overall risk score is a function of different driving conditions, such as crash, near miss and normal driving. One can further increase the model resolution by creating subcategories of each. For instance, there may be several subcategories of normal driving depending on safety critical applications. One can also combine several risk categories together and reduce driving states to "unsafe" and "normal". Here we assumed three categories of 'crash', 'near miss' and 'normal driving'. Our methodology includes the estimation of each of these probabilities and a function that combines them into a risk score. In the absence of sufficient data one may use Probability of crash as a risk score. With the availability of data from drivers one will be able to update the model parameters and also extend the model to include additional data variables. Therefore, risk score modeling is a dynamic process that will improve as more data becomes available. Given a sample of real-time driving events one can train a Generalized Linear Model (GLM) with regularization (elastic-net) and estimate the regression coefficients.

• Preliminary architecture for an on-board driver alert system

Our database supports the risk data collected from the above prediction model. The risk score model can be programmed offline and run from a cloud based server at sampled times. It can also be programmed by an APP and run from a smartphone inside vehicle. We have adopted a two-tried approach. A full prediction model estimation is programmed using R in an offline cloud based setting. The model parameter estimates are then programmed using a tabular structure in an APP. This APP also collects driver information (ID, age, etc.) from the database. A driver ID and GPS origin and destination define a trip. More complex trip models can be built at some later phases. The APP starts at the GPS origin. At intervals defined by the driver (e.g., default can be set to every 10 to 15 seconds) the risk score value (using either predictive or rule based approach) is calculated and stored locally in the smartphone. The sampled valued are also graphed (y-axis is risk and X-axis is the distance travelled). We refer to these graphs by driver risk profile over a trip. These risk profiles are uploaded to a server at the completion of the trip along with other pertinent information (such as some important driving patterns, etc.).

FINDINGS

Partially supported by this funding, we developed a novel data-driven approach to real-time traffic safety risk prediction. We quantified traffic safety risk as the likelihood of adverse driving outcomes. We proposed using the elastic net regularized multinomial logistic regression with a built-in variable selection and shrinkage mechanism and a cost-sensitive loss function for imbalanced data. To the best of our knowledge, this work is the first attempt toward applying the elastic net model to traffic accident related research. We introduced five measures of goodness (lack of goodness) to evaluate the performance of the prediction models, namely miss-classification, Type I, Type II, Upper Off-diagonal Triangular and Lower Off-diagonal Triangular error rates to take into account the sensitivity and specificity of the classifier in identifying cases in the minority classes. To evaluate the prediction performance of the prediction models, we used 10-fold cross validation, and used a subset of SHRP 2 NDS safety data to show the applicability of our proposed approach. We also developed a data preparation and feature engineering approach, which included the following steps. First, we introduced a three-class, i.e. crash, near-crash, and normal-driving, prediction model. Then, by breaking down the normal-driving class, we further illustrated that the model can easily accommodate different resolutions of driving outcome.

The developed predictive models can be used to support a multitude of applications, including Advanced Driver Assistance System (ADAS), safety risk profiling of drivers, and safety risk scoring of roadway segments for dynamic hotspot analysis. For example, the models can be incorporated into a data-driven collision warning system to warn drivers of critical events and/or unsafe driving situations. The design of such a system and its practical issues, including sensors acquisition, in-vehicle perception, onboard processing and the HMI design, can be the subject of future works.

In this study, in selecting the final models, we preferred a higher type-I error in exchange for a lower type-II error since the cost of the latter can be loss of lives. Since there is a tradeoff between type I and II errors, the effort to reduce one generally results in increasing the other. One approach to simultaneously reduce type I and II errors in an imbalanced data setting is to collect more data to increase the sample sizes of minority classes. In fact, in this study we have only used one third of the SHRP 2 data. Including more crash and near-crash cases most likely improve the prediction performance of the models. Another approach to tackle this problem is to use re-sampling methods such as bootstrapping which can be the subject of a future work. Additionally, with more data and provided that there are enough samples per specific crash type, separate models can be developed to predict specific types of collisions. This, in turn, will raise the issue of how to prioritize different crash-type warnings at a given time in an ADAS. Finally, advanced machine learning algorithms such as deep learning can be a potential alternative to improve the performance of the safety risk model as big players like Google, Nvidia, Microsoft and IBM have already invested heavily in new projects around this breakthrough method.

The funding from this project partially supported a PhD thesis, which was completed and defended on January 17, 2017. A technical article was also accepted and schedule for publication:

"Traffic Safety Risk Prediction Using Driver Behavior and Roadway Information Data" Nasim Arbabzadeh and Mohsen A Jafari, to appear in *IEEE Transactions on Intelligent Transportation Systems*, 2017.

We also built a preliminary version of a smartphone APP was tested in limited basis by our team. Additional work will be necessary to build the alpha version of the APP and run additional field testing.

CONCLUSIONS

Future roadways will have a mix of autonomous and automated vehicles with regular vehicles that require human operators. To ensure the safety of all the road users in such a network, it is necessary to enhance the performance of the present Advanced Driver Assistance System (ADAS) for lower classes of vehicles. Real-time driving safety risk prediction is an essential element of an ADAS. In this study, we developed a novel data-driven approach to predict traffic safety risk that can be customized to individual drivers by including driver-specific variables. In particular, we used the elastic net regularized multinomial logistic regression and data from the second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study (NDS) to build these predictive models. We rigorously examined the variables in the dataset and performed data preparation and feature engineering steps to enhance the prediction performance with respect to model predictors. Two versions of the model were developed according to the level of warnings that the model can generate based on driving conditions.