An Improved Accident Crash Risk Prediction Model Based On Driving Outcomes Using Ensemble Of Prediction Algorithms

Ybralem Bugusa, Shruti Patil

Abstract: Safety and protecting our self from accidents that causes sudden life lose, injury and damage is the wish of every person. To bring this safety it is crucial to identify the main factors that cause the accident and find solutions. Prediction of real-time risk is among the solutions that aware the drivers to concentrate on their driving during driving process. Real-time risk prediction is important part of Advanced Driver Assistant System (ADAS). The main objective of this study is developing model that predict the driving risk during driving situation. We have used the Virginia Tech Transportation Institute (VTTI) data set with three events crash, near-crash and normal state. 16 variables with 15 independent variables are considered which we consider them as relevant variable in crash risk prediction and 1 dependent variable. The variables are selected from driver information, roadway information and weather condition. For this investigation, we compared the result of Elastic net and individual algorithm in the ensemble with the ensemble model. Resampling with replacement to improve the accuracy of minority class. According our experiment ensemble algorithm performs better overall accuracy than Elastic net and other individual algorithms used as base learner in ensemble model.

Index Terms: Traffic safety, VTTI dataset, Ensemble algorithm, Elastic net, resampling with replacement.

1. INTRODUCTION

Traffic accident is one of the top reasons that takes away human life and cause damages in world-wide [17][22]. The Global status report on road safety 2015, collected data from 180 countries states that 1.25 million road traffic deaths are happened per year. And reported with high accident rate in low-income countries. Many countries state their law and control on using seatbelt use, drink-driving and speed which is better for life safety. To overcome such human life lose, moral and infrastructure damage, it is necessary to assess who is responsible and what are the solutions. Many data collection method and analysis are done to bring this solution. Currently, with growing of technology advanced technological devices i.e. cameras, sensors, loop detector and more are used to observe factors that cause this accidents in many direction such as driver and vehicle information, weather condition and type of accident or event occurred. Many investigations are relied on data collected from these devices to build models that predicts the level of crash and explore the responsible one from the factors studied. Moreover, studies focus on identifying the factors that affect the target variable or driving outcome. The finding of various studies considers that driver behaviour is as potential accident risk factor [1][12]. And also the secondary tasks such as driver distraction and inattention, are the main factors in traffic accident investigation [22-24].

- *Ybralem Bugusa is a student at Symbiosis International (Deemed) University, Pune, India. ybralem.wekele@sitpune.edu.in*
- *Shruti Patil is Assistant Professor at Department of Computer Science and Information Technology, Symbiosis International (Deemed) University, Pune, India. shruti.patil@sitpune.edu.in*

Additionally, studies revealed that the cause of the crash and near-crash is with violation of traffic rules such as seatbelts, driving drank are the most significant factors influencing injury severity [17]. They also find that driver age, gender, driving experience and road type are potential factor in crash risk accident [17][18][20]. Decision tree algorithm can identify and simply explain the association between variables and crash risk and do not need to provide an argument [18]. It finds out that, incorrect overtaking and not using a seatbelt are the most significant factors disturbing the severity of injuries. Authors [25] of work used A negative binomial regression model to identify factors caused individual driver risk and K-mean cluster algorithm for classification and prediction of level of risk. Their investigation result that the critical incident rate is an effective predictor for high-risk drivers and drivers age, personality, and critical incident rate had insignificant impacts on crash and near-crash accidents. According to [26] Work, driving anger, impulsiveness and aggressiveness are the main predictors of aggressive and transgressive behaviors on the street. The finding indicates that transgressive driving behaviors are significant pointers of aggressive driving. To develop this they have used multiple regression analysis methodology. [5] Study compared four algorithms i.e. Multinomial Logit, Nearest Neighbor Classification, Support Vector Machines and Random Forests to predict crash severity. It used 2012–2015 reported crash data from Nebraska data set and Nearest Neighbor Classification got better result of the remaining three. Authors of [15] compared Gradient Boosting and Generalized Linear Model for prediction of insurance cost loss and finds Gradient Boosting performs better accuracy. Moreover, currently instead of using a single learner using multiple diverse or the same algorithm, then combining the result with another machine learning methodology attracts the attention of many researchers. This method is called Ensemble algorithm, and it is applied on many different areas of study to improve a decision system's strength and accuracy of single machine learning techniques [4][28]. Study of [11] conducted ensemble algorithm with diverse algorithms i.e. Artificial Neural Networks, Naïve Bayes, and Decision Tree and generalize their result using genetic algorithm (GA) with weighted-averaging method to derailment accident prediction. Their finding states that ensemble algorithm got

better predictive result than individual algorithms. Authors [13] develops A novel ensemble artificial neural networks (EANN) model by comparing result with regression analysis for brain death prediction. And specified that ensemble model got better performance even in complicated problem as brain death prediction. Investigation on improving Student Enrollment Prediction [2] proves that ensemble algorithm performs good accuracy than using single algorithm. Also they stated that, this mechanism can provide better rules for considering the factors that influence student enrollment. The aim of this paper is improving prediction of driving outcome by considering factors potentially influence accident real time crash/near-crash and that expose to minimum risk using ensemble algorithm. It uses Virginia Tech Transportation Institute (VTTI) dataset [20-21]. The predictive accuracy of the model is compared against the Elastic net and various algorithms. Overall flow of the work is shown below.

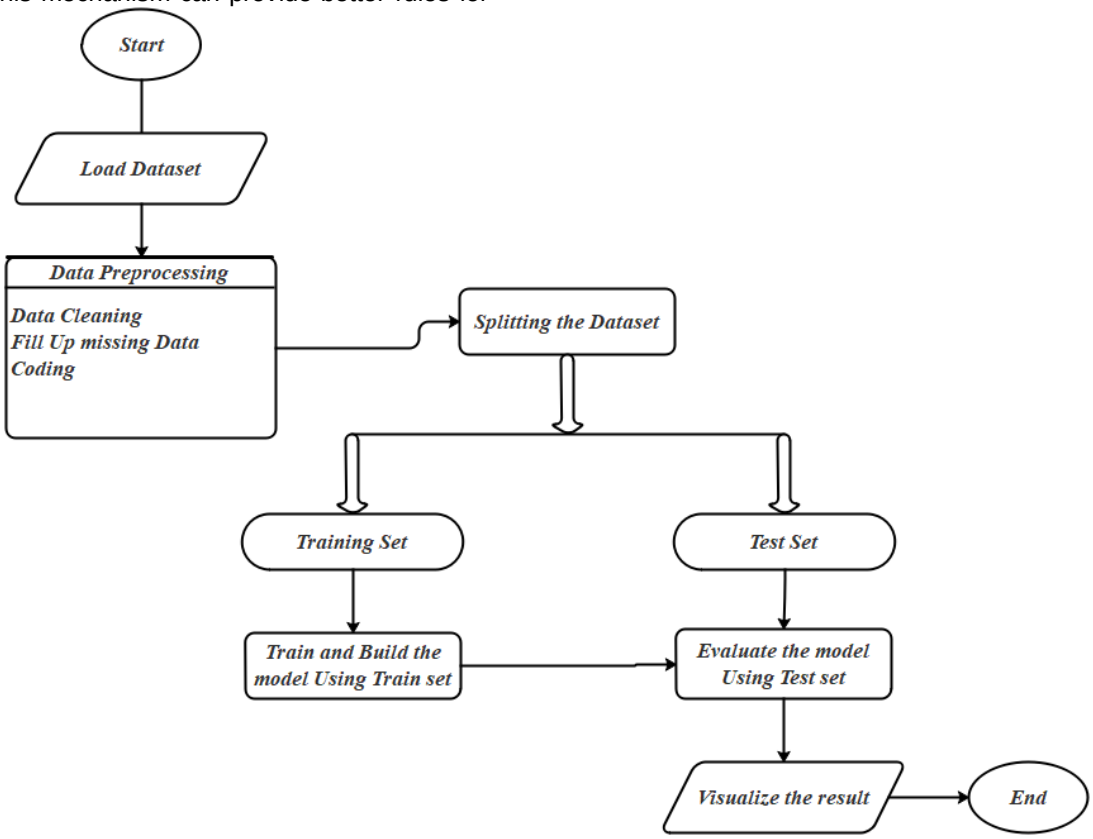

Fig 1: *Overall flow diagram*

The remaining section summarizes a review dataset preparation, methodology used in our work and how it works with computation of the variable importance used, the result and challenges, and finally, the conclusion and future work is mentioned.

2. DATA SET

2.1 Data source

In this study, we have used Virginia Tech Transportation Institute (VTTI) Naturalistic Driving Study which is the free data set of Secondary Strategic Highway Research Program (SHRP2) Naturalistic Driving Study. It collects 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. And includes equipping volunteer participants' vehicles with advanced, unobtrusive instrumentation (e.g., cameras, sensors, radar) that automatically and constantly collects driving parameters—including speed, time to collision, global positioning system (GPS) location, acceleration, and eye glance behavior[16][24][21][30-31].
The data set we have got consists of 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.

2.2. Data preparation and description

To make the data easily understandable, we have further pre-process it using Microsoft Azure Machine Learning Studio. It is a collaborative, drag-and-drop tool you used to build, test, and deploy predictive analytics solutions on your data [32]. In pre-processing data cleaning, punctuation correction and missed value are handled. Using this tool we have remove the rows that contains no value. Since we are working with highly imbalanced data set. Random sampling

technique is applied on major class (i.e. baseline event data) to reduce the data size required to achieve better classification. After data cleaning and reduction, 3353 total data are left which consists of 2595-Baseline (B), 969-NearCrash (N) and 62- Crash (C) Event Datasets with selected 16 important variables which includes driver state, and driving environment information. For more safety, especially for drivers with limited skill the baseline (normal) event is further decomposed in to three class or event, based on the six variables Driver Behaviour 1, 2, 3 and Distraction 1, 2, 3. Then we have again clean and merge the classes.

TABLE 1:
Variable description

No	Variable	Description	Group	Variable Type
1	Age	Subject age on entry to study.	Driver	Categorical
2	Gender	Subject gender.	Driver	Binomial
3	Driver behavior	Behaviours resulting from the context of the driving environment that include what the driver did to cause or avoid the crash or near-crash.	Driver	Categorical
4	Distraction	driver engagement in secondary tasks	Driver	Categorical
5	Surface condition	Type of roadway surface condition that would affect the vehicle's coefficient of friction.	Roadway-information	Categorical
6	Traffic flow	Roadway design (including the presence or lack of a median)	Surrounding externality	Categorical
7	Traffic lanes	Number of lanes the subject vehicle could easily maneuver into, including any turn lanes, acceleration lanes, etc., not taking into account any occupants of these lanes	Roadway-information	Integer
8	Traffic density	Based entirely on number of vehicles, and the ability of the driver to select the driving speed	Surrounding externality	Categorical
9	Traffic control	Type of traffic control applicable to the vehicle at the time of the event	Roadway-information	Categorical
10	Relation to junction	Point where 2 roads meet	Roadway-information	Categorical
11	Alignment	Geographical description of the roadway	Roadway-information	Categorical
12	Locality	Best description of the surroundings at the time of the start of the event. If there are ANY commercial buildings, indicate as business/industrial area	Roadway-information	Categorical
13	Lighting	Lighting condition at the time of the event.	Environmental condition	Categorical

14	Weather	Weather condition at the time of the event.	Environmental condition	Categorical
15	Driver seatbelt use	Driver's use of seatbelt at the time of the event.	Driver	Categorical

After decomposing and cleaning we have got the following five class data distribution. The colours for each class represent the level of accident risk.

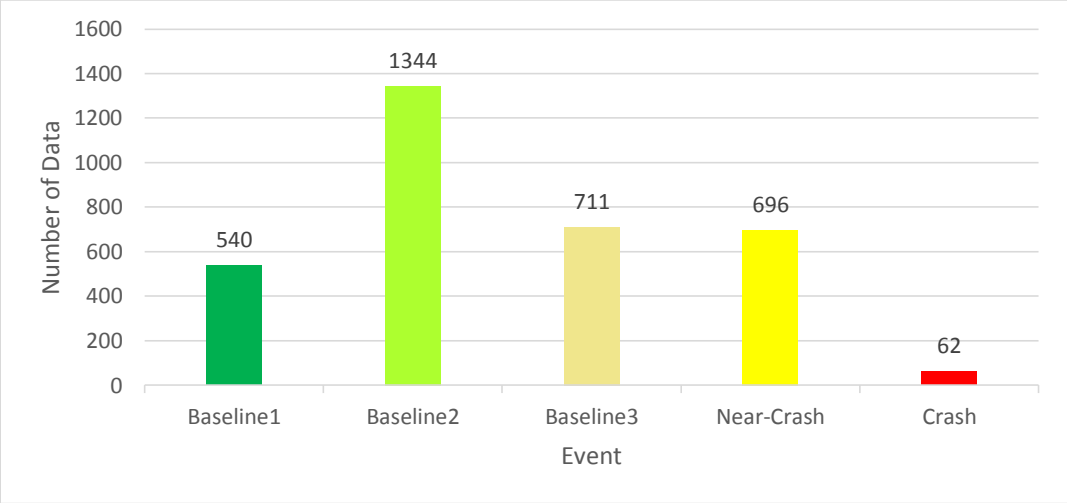

Fig2: Distribution of dataset in all outcomes.

3. METHODOLOGY

For our experiment we have compared the result of Elastic net and Ensemble algorithm to get better prediction accuracy using R tool

3.1 Elastic net model

Elastic net is supervised classification model used in many areas for variable selection, optimization and classification purpose. It is hybrid of Lasso and Ridge multinomial logistic regression that have the ability to handle bias and variance trade-off, highly correlated variables and work with categorical and numeric data[6-9].

3.2 Ensemble model

Ensemble model is a machine learning model that used more than one or multiple algorithm, which is called base learner and generalise their result with other algorithm for prediction and analysis on multi discipline. There different methods of ensemble model:

a. Bagging

Bagging or Bootstrap Aggregation is the technique that combines multiple algorithm with bootstrap sample and replacement. This used for improving prediction accuracy of individual algorithm.

E.g. Random Forest algorithm.

b. Boosting

Boosting it combines multiple learner algorithms with overweight techniques it is used to reduce bias.

E.g. AdaBoost algorithm

c. Stacking

Staking it is also known as stacked generalization includes a learning algorithm to generalize the predictions of diverse multiple other algorithms.

Firstly, it trains the individual algorithm used in developing the model. Then other algorithm or it can be among the trained previous used as combiner by using the single learner algorithm as input or independent variable to predict the target class. It uses logistic regression method to combine the trainers. In our work we have used the stack algorithm of C5.0, K-Nearest Neighbor, J48 and Naïve Bayes as base learner and Gradient Boosting Machine (GBM) as combiner or Meta classifier to predict the target class. The Meta classifier used the predicted value of the base learners to generalize the result and for precise decision.

Algorithm of stack ensemble

Input: Data set D = $\{\mathbf{xi}, \mathbf{yi}\}_{\mathbf{i=1}}^{\mathbf{m}}$

Base learner algorithms: $\mathbf{L_1}, \ldots \mathbf{L_T}$

Top level learning algorithm: $\mathbf{L_S}$

Output: Ensemble Classifier H

step1: Train a First individual learner $\mathbf{h_t}$ by applying the first-level

for t = 1 to T

$h_t = L_t(D)$

End

step2: Generate new dataset from base learner algorithm result

$D_1 = \{\}$

for i = 1 to m

for t = 1 to T

$Z_{it} = h_t(xi)$

End for

$D_1=(Z_{it}, y_i)$

End for

step3: Train h' based on new data set $\mathbf{D_1}$

$h' = L_s(D_1)$.

$H(x) = h'(h_1(x); \ldots; h_T(x))$

4. RESULT AND DISCUSSION

As it is mentioned in previous section, we have compared elastic net as well as individual algorithms used in stacking as base learner with Meta classifier. The overall flow of our work look like this: To build model with elastic net algorithm it uses glmnet library which uses the hyper parameter alpha value lies between 0 and 1. It can be tuned using cross-validation to compute lambda value to handle model optimization. The proposed algorithm will search the optimum lambda value with minimum cross validation error (cvm) by taking nine (9) alpha values between 0.1 and 0.9. In order to calculate lambda value 10 fold cross-validation applied on the model. The sample with replacement is used to partition our dataset in to training and testing sets. This sample technique increases the dataset from 3353 to 4200 by taking 847 data randomly from original dataset and replace it. This increase the prediction probability of minority class. The data is split in to 75% training set and 25% testing set. The foreach used in R tool to search the lambda value using 10-fold cross validation and performs parallel computing.

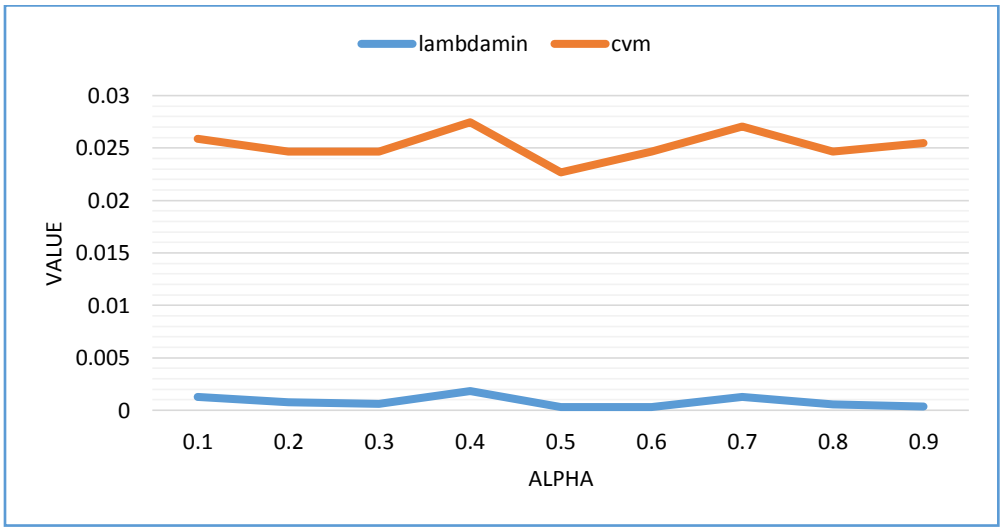

Fig 3: *Alpha value Vs Lambda value to get optimal values*

The experimental result indications states that the optimal lambda value is 0.0002765947 at alpha value 0.5, and confusion matrix shows that the model achieves overall 95.22% accuracy and kappa 93.42% result.

TABLE 2:
Elastic net confusion matrix

		Actual				
		BS1	BS2	BS3	Near-Crash	Crash
Predicted	BS1	263	0	0	5	1
	BS2	0	629	1	4	1
	BS2	0	0	321	8	1
	Near-Crash	2	2	8	297	22
	Crash	2	2	4	13	5

In addition to the model developed using elastic net, we have conducted ensemble model which combines multiple algorithms and generalize their prediction result to achieve better accuracy. This idea relates with information fusion in the area of Internet of Things. Information's are gathered from different sensors from different sources then the information are integrated and filter relevant information which is helpful for future decision. The ensemble model includes the combination of base learners: C5.0, K-Nearest Neighbor, J48 and Naïve Bayes as well as Gradient Boosting Machine (GBM) as super learner. The dataset are split in to 75% training dataset and 25% test dataset. Repeated 10-fold cross validation are used to train the individual algorithm in the ensemble model using the training dataset. To change the candidate values of the tuning parameter. The train function can generate a candidate set of parameter values and the tune-Length argument controls how many are evaluated. We have used 3 tune-Length and evaluate integers between 1 and 3. This allows to realize the effect of model tuning parameters on performance this helps for choosing the optimal parameters used for model building. The sample with replacement is true which slightly increase the accuracy of minority class. This takes some portion of data from original dataset and replace it. This may increase

the size of minority class to exist. The overall process can be visibly shown the ensemble algorithm.

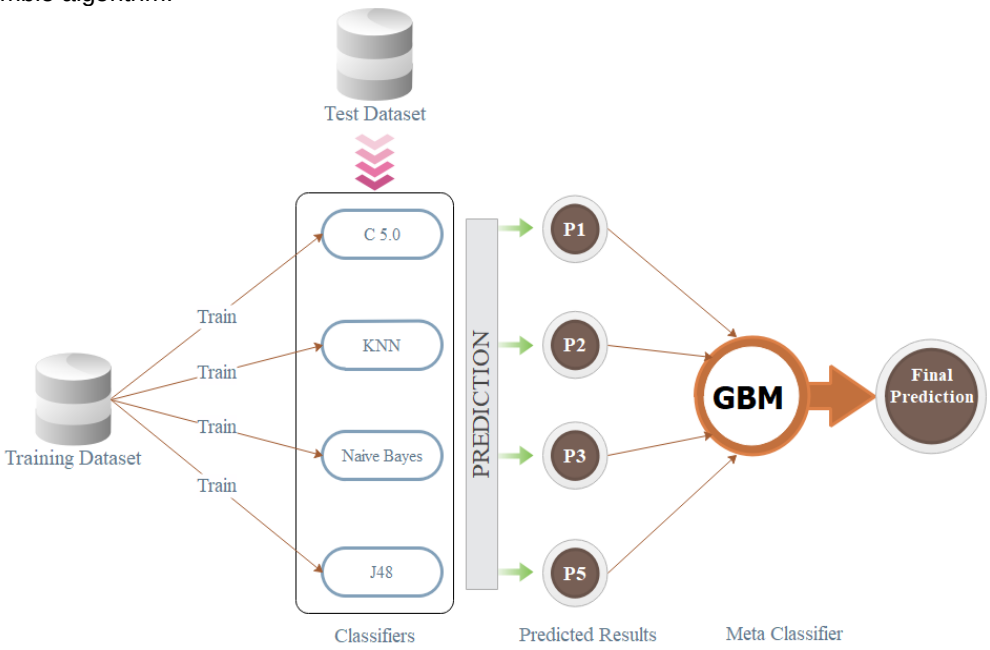

Fig 4: *proposed ensemble architecture*

As we are looking in the above architecture after training the model and testing with the test dataset to observe the performance of the algorithms in unseen data, we have generated new dataset by saving the predicted value which we consider Meta data of the individual models for training as well as for testing in different frame. The Meta classifier uses the new generated dataset to predict the target class which is the driving outcome. This is what we call the stack ensemble model. We got the accuracy and kappa metric of the ensemble model using confusion matrix.

TABLE 3:
Ensemble model confusion matrix

		Actual				
		BS1	BS2	BS3	Near-Crash	Crash
Predicted	BS1	271	0	0	5	2
	BS2	0	655	0	3	2
	BS2	0	0	332	5	2
	Near-Crash	0	0	11	300	18
	Crash	0	0	0	3	3

After the experimental analysis the ensemble algorithm achieves better accuracy (96.84%) and kappa (95.6%) in comparing with individual algorithm in the ensemble as well as with elastic net. At confidence level of (95%)

The summary of resample results Comparison of algorithms by using 10 fold validation applying on the whole models used in ensemble algorithms shown as below. It used to calculate the Accuracy and Kappa metrics of the individual algorithm in the ensemble model and the ensemble model it self

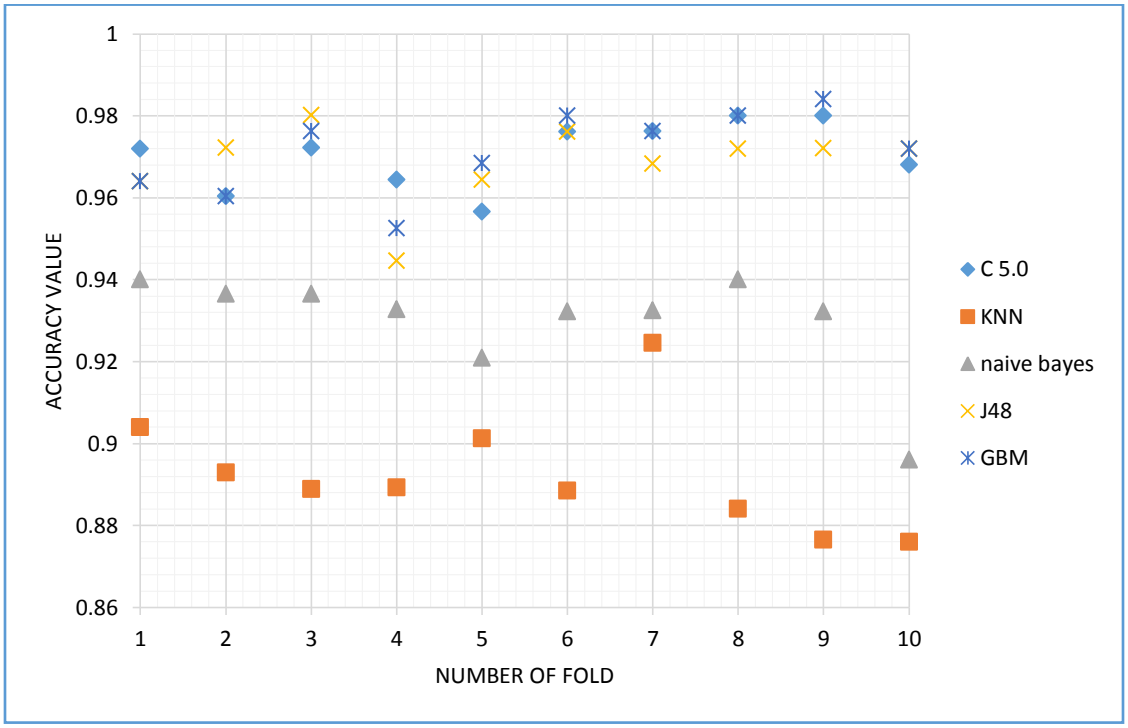

Fig5:10-fold resample accuracy comparison of the models

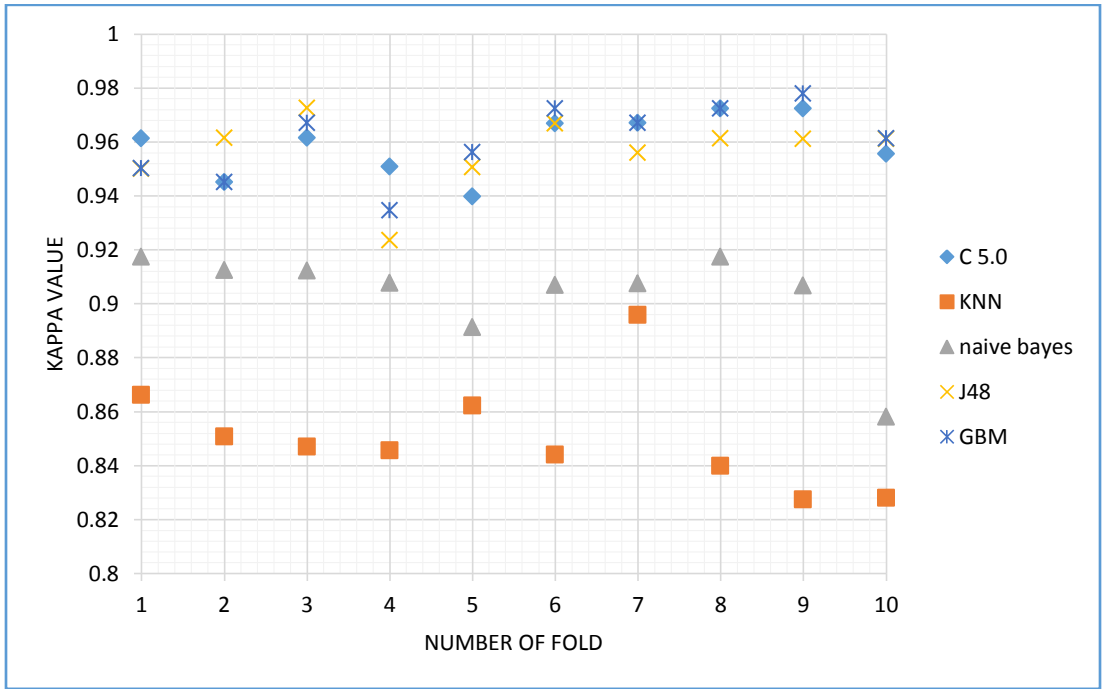

***Fig 6: 10**-fold resample Kappa's result comparison of the models*

We have also investigated the importance of variables used in the models to predict the target class. According to the analysis, Driver behavior and secondary task involvement (Distraction) and subject age can highly use in deciding of the accident risk level. The importance of variables used are ranked based on information gain they provide to the model in prediction.

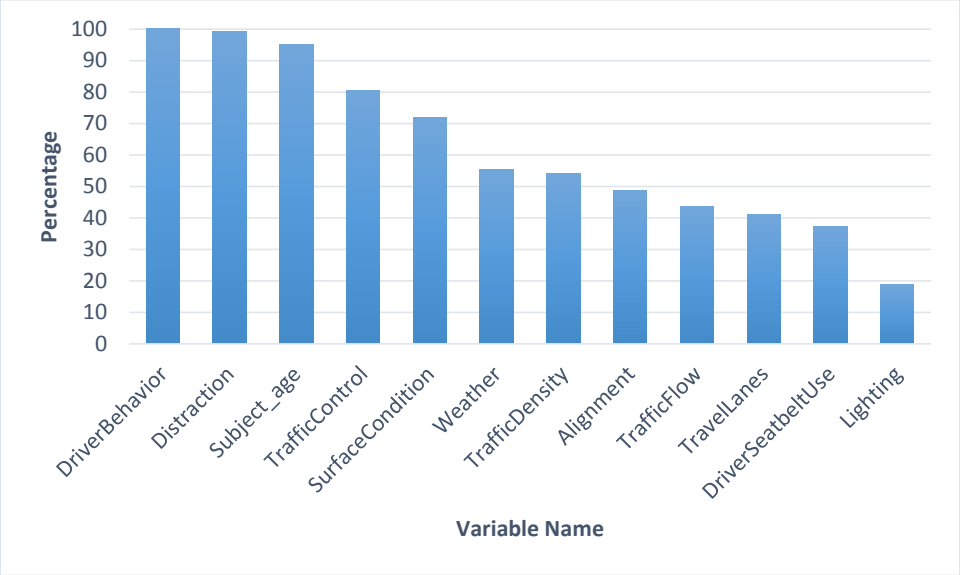

Fig 7: *Variable importance list*

The result of experiment shows that there is more misclassification error in minority class. Such problem can be reduced by considering more training data or using bootstrap sampling technique. Using bootstrap can take more computing time and space.

5. CONCLUSION

In this study we have offered the stack ensemble methodology of C5.0, K-Nearest Neighbor, J48 and Naïve Bayes as well as Gradient Boosting Machine (GBM) as super learner. To predict the target class. The ensemble model achieves better accuracy comparing with base learner as well as with Elastic net. Sampling with replacement increases the prediction of driving outcome. The variables used to feed the models are selected from driver information, roadway information and weather condition. Considering vehicle information, time series data with respect of driver, environment condition and specific variable of driver behaviour for crash and near-crash events can be further improvements for this study. In addition, using advanced machine learning like deep learning approaches can improve the performance of the system.

REFERENCES

[1] Arbabzadeh, Nasim, and Mohsen Jafari. "A Data-Driven Approach for Driving Safety Risk Prediction Using Driver Behavior and Roadway Information Data." IEEE Transactions on Intelligent Transportation Systems (2017).

[2] Wanjau, Stephen Kahara, and Geoffrey Muketha. "Improving Student Enrollment Prediction Using Ensemble Classifiers." (2018).

[3] Sameen, Maher Ibrahim, and Biswajeet Pradhan. "Severity Prediction of Traffic Accidents with Recurrent Neural Networks." Applied Sciences 7.6 (2017): 476.

[4] Zhang, Cha, and Yunqian Ma, eds. Ensemble machine learning: methods and applications. Springer Science & Business Media, 2012.

[5] Iranitalab, Amirfarrokh, and Aemal Khattak. "Comparison of four statistical and machine learning methods for crash severity prediction." Accident Analysis & Prevention 108 (2017): 27-36.

[6] Li, Qiang, et al. "Correlated logistic model with elastic net regularization for multilabel image classification." IEEE Transactions on Image Processing 25.8 (2016): 3801-3813.

[7] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent." Journal of statistical software 33.1 (2010): 1.

[8] Boulares, Mehrez, and Mohamed Jemni. "Learning sign language machine translation based on elastic net regularization and latent semantic analysis." Artificial Intelligence Review 46.2 (2016): 145-166.

[9] Thanathamathee, Putthiporn, and Chidchanok Lursinsap. "Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques." Pattern Recognition Letters 34.12 (2013): 1339-1347.

[10] Polikar, Robi. "Ensemble learning." Ensemble machine learning. Springer, Boston, MA, 2012. 1-34.

[11] Kaeeni, Samira, Madjid Khalilian, and Javad Mohammadzadeh. "Derailment accident risk assessment based on ensemble classification method." Safety Science (2017).

[12] Gregoriades, Andreas, and Kyriacos C. Mouskos. "Black spots identification through a Bayesian Networks quantification of accident risk index." Transportation Research part C: emerging technologies 28 (2013): 28-43.

[13] Liu, Quan, et al. "Brain death prediction based on ensembled artificial neural networks in neurosurgical intensive care unit." Journal of the Taiwan Institute of Chemical Engineers 42.1 (2011): 97-107.

[14] Abellán, Joaquín, Griselda López, and Juan De OñA. "Analysis of traffic accident severity using decision rules via decision trees." Expert Systems with Applications 40.15 (2013): 6047-6054.

[15] Guelman, Leo. "Gradient boosting trees for auto insurance loss cost modeling and prediction." Expert Systems with Applications 39.3 (2012): 3659-3667.

[16] Neale, Vicki L., et al. "An overview of the 100-car naturalistic study and findings." National Highway Traffic Safety Administration, Paper 05-0400 (2005).
[17] Zhang, Guangnan, Kelvin KW Yau, and Guanghan Chen. "Risk factors associated with traffic violations and accident severity in China." Accident Analysis & Prevention 59 (2013): 18-25.
[18] Kashani, Ali Tavakoli, and Afshin Shariat Mohaymany. "Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models." Safety Science 49.10 (2011): 1314-1320.
[19] Wright, Nicholas, and La-Troy Lee. "New Evidence on the Causal Impact of Traffic Safety Laws on Drunk Driving and Traffic Fatalities." (2017).
[20] McClafferty, Julie, and Jonathan M. Hankey. 100-car reanalysis: Summary of primary and secondary driver characteristics. Virginia Tech. Virginia Tech Transportation Institute, 2010.
[21] Tian, Renran, et al. "Studying the effects of driver distraction and traffic density on the probability of crash and near-crash events in naturalistic driving environment." IEEE Transactions on Intelligent Transportation Systems 14.3 (2013): 1547-1555.
[22] Klauer, Sheila G., et al. "Distracted driving and risk of road crashes among novice and experienced drivers." New England journal of medicine 370.1 (2014): 54-59.
[23] Klauer, Sheila G., et al. An analysis of driver inattention using a case-crossover approach on 100-car data. No. HS-811 334. 2010.
[24] Klauer, Sheila G., et al. "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data." (2006).
[25] Guo, Feng, and Youjia Fang. "Individual driver risk assessment using naturalistic driving data." Accident Analysis & Prevention 61 (2013): 3-9.
[26] Berdoulat, Emilie, David Vavassori, and María Teresa Muñoz Sastre. "Driving anger, emotional and instrumental aggressiveness, and impulsiveness in the prediction of aggressive and transgressive driving." Accident Analysis & Prevention 50 (2013): 758-767.
[27] Beshah, Tibebe, and Shawndra Hill. "Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia." AAAI Spring Symposium: Artificial Intelligence for Development. 2010.
[28] Dietterich, Thomas G. "Ensemble methods in machine learning." International workshop on multiple classifier systems. Springer, Berlin, Heidelberg, 2000.
[29] Behnood, Ali, and Fred L. Mannering. "The effects of drug and alcohol consumption on driver injury severities in single-vehicle crashes." Traffic injury prevention 18.5 (2017): 456-462.
[30] Brewer, Marcus A., Shannon Barkwell, Pei-Sung Lin, Seckin Ozkul, Walter R. Boot, Priyanka Alluri, Larry T. Hagen et al. "Exploration of the SHRP2 naturalistic driving study data to identify factors related to the selection of freeway ramp design speed." Ann Arbor 1001 (2017): 48109-2150.
[31] Dingus, Thomas A., Feng Guo, Suzie Lee, Jonathan F. Antin, Miguel Perez, Mindy Buchanan-King, and Jonathan Hankey. "Driver crash risk factors and prevalence evaluation using naturalistic driving data." Proceedings of the National Academy of Sciences 113, no. 10 (2016): 2636-2641.
[32] Docs.microsoft.com. What is Azure Machine Learning Studio?. [online] Available at: https://docs.microsoft.com/en-us/azure/machine-learning/studio/what-is-ml-studio [Accessed 18 December. 2017].