Traffic Data Analysis in Residential and Urban Areas based on Machine Learning Techniques

Madhuri Dadhich

Department of Computer Science
Christ University Faculty Of Engineering
Bangalore, India

madhuri.dadhich@mtech.christuniversity.in

***Abstract*—With the successive increase in usage of vehicles, severe traffic congestion is on the rise. This in turn leads to increase in environmental pollution and accidents which ultimately affects the safety, time consumed and money spent of the transport users. .Road safety could be enhanced by decreasing the traffic crashes. Traffic crashes cause traffic congestion as well, which has become unbearable, especially in mega-cities .This paper uses machine learning algorithms codes with python to predict the crashes and driver's behaviour.This paper analyses (1) Traffic accidents dataset of 146322 examples, find useful insight and patterns from the data, and forecast possible accidents in advance (2) Data analysis on a vehicular casualties dataset of 194477 examples, to study the driver's behavior on the road .**

***Keywords*— Road Safety, Traffic Accidents,Data Analysis,Machine Learning**

I. INTRODUCTION

The world has billion of vehicles on the roads nowadays and people have become more and more dependent on transportation. This increasing number of transportation vehicles will lead to more incidents, more congestion, more loss of time, money, and more harm to the environment around us. The world should work collaboratively, academia and industry to find solutions for the aforementioned problems that could happen from the increasing number of vehicles on the roads

The analysis of the large-scale data of transportation and accidents has many potentials and it can give very useful insights from the hidden relations in the data .Accidents dataset which consists of 146322 examples and the Casualties dataset, which consists of 194477 examples. We selected these two datasets because they represent real-time data. We trust using these data because they came out from actual accidents on public roads that were reported to the police, and subsequently recorded, using the STATS19 accident reporting form [2]. The advantage obtained from using the first dataset is to find out the main causes of traffic accidents, which mainly, causes traffic casualties and congestion, and we aim to prevent both of them. The casualties dataset used to study the human behavior effect on causing traffic accidents. Human actions on vehicles or roads would significantly cause different side effects. Human behavioral actions should be studied thoroughly, because they have significant impacts on traffic and roadways.

II. RELATED WORK

Krishnaveni and Hemalatha [5] utilized the Hong Kong's Transportation Department accidents data for the year of 2008, with a total number of 34000 records. To obtain the causes of accidents, a number of classification algorithms were used and their performance compared in WEKA. The results presented that Random Forest outperformed J48, NB, PART and AdaBoostM1 classifiers.Chong et al. [6] used the National Automotive Sampling Systems Data (NASS) and General Estimates Systems Data (GES) to study traffic accidents injury severity .They built a machine learning models to classify the severity of five categories of injuries, in order to find the relationship between the factors of the drivers, vehicles and roadways, and the injury severity. They have used ANN, Decision Trees, SVM and a hybrid approach combining Decision Trees with ANN. They only exploited 10500 records in their experiments, which is not enough to study all hidden and potential relationships for such a problem. Finally, their approach presented that the main cause of fatal injury, is speeding over the lawful limit. Florido et al., [2] developed an application of patterns and behavioral models of time series for the data collected by sensors belonging to the Spanish Directorate General for Traffic (SDGT). The application predicts the behavior of the system to get an early notification of traffic congestion in order to give an alternative circulation of vehicles. Authors of this paper have used a total number of examples around 36000 records for both testing and training, with eleven attributes, eight quality parameters and three classification algorithms C4.5, ANN and NN. All of these models have been built and experimented with KEEL mining tool [5]. Results showed that C4.5 outperformed both ANN and NN. Hamdy and Behrouz [3] have studied the challenges associated with transportation systems. Authors introduced a design only of a real-time transportation data mining framework. The theoretically proposed framework core idea is based on analyzing real-time traffic data in conjunction with decision support systems and makers. In their paper, road accidents' data from UK Department for Transportation (DFT) for the period of 2005-2012 has been used. Due to the tools and resources limitations, the authors of this paper utilized a small portion from the data of 5486 records for both training and validation phases. Despite this research gave a good analysis for the traffic crashes from the used data, but using the full datasets in our paper gave different results. This pager utilizes the full

traffic data set and uses machine learning algorithms to obtain better insight and more accurate results.

III. EXPERIMENT

Machine learning is a burgeoning technology for mining knowledge from data, where a lot of experts are taking it seriously to solve big problems of interest [10]. Too many learning algorithms, methods, tools, and techniques have been implemented in computer programs to easily apply learning methods on a dataset and analyze it to get useful insights and relationships. There are traditional machine learning tools, such as R and python but these conventional tools cannot handle and scale to Big Datasets..In this work, we have used python and machine learning techniques design and run our experiments .This paper uses two machine learning algorithms :Naive Baye's and Random Forest for feature selection and data analysis .In this work, we study two workbench big datasets, the Accidents dataset which consists of 146322 examples and the Casualties dataset, which consists of 194477 examples. Table 1 gives the description of the data sets that has been used for the experiment..This paper uses 80% data for the training and remaining 20% data for the testing purposes

We have selected the accident severity attribute as the testing value (class) for the two datasets in the experiment .Our datasets now have the class of severity (Slight and Fatal). This conversion into a binary class helped in classifying the crash prediction.

Although we have many attributes in each dataset, this does not mean that all of them are necessarily needed and they will give better results. Decreasing the number of attributes (features) will decrease the processing time and increase the prediction accuracy .This paper uses Naive Bayes as the classifier to select the features in our proposed approach in addition to its fast computation time. After applying feature selection on both datasets, we have selected 9 attributes for the Accidents dataset and 8 attributes for the Casualties dataset. Results show that feature selection significantly decreases the computation time.

TABLE I

DATASETS DESCRIPTION

Dataset Name	#Records	#Attributes
Accidents_2014	146322	32
Casualties_2014	194477	15

III. RESULTS AND ANALYSIS

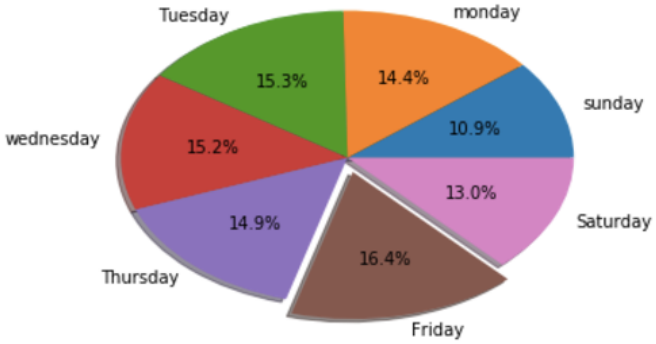
From the analysis of the accident's data and by discovering the hidden relationships, we can extract the main causes of accidents that lead mainly to traffic congestions. Each accident will be classified either as Fatal or Slight, thus our analysis revealed the following:

1) 64.9% of all the accidents happen on residential areas where the speed limit is 30 kmh. Most of the accidents don't happen on highways as most people think.
2) 65.8% of all the accidents occur in Urban areas
3) In uncontrolled junction or a give way junction 50% of the accidents take place there, with majority of slight accidents.
4) People would think that most of the accidents take place in bad weather, especially when there is snow or fog. Our analysis shows that 81% of the accidents happens when the weather is fine and has no winds
5) The same issue with the weather condition, people might think that most accidents occur when there is no light or at night, but analysis showed that 73.8% of the total number of accidents happened in daylight
6) On Fridays, when the weekend starts, policies and speed limits should be changed in Urban areas and single carriage ways because Fridays have the highest percentage of accidents of 16%. Figure 1 shows the accident percentage of everyday of week

In this work, by studying the human behavior and impact, we can design new rules to the drivers depending on their age, sex, education level, marital status, and many more individual or combined characteristics. We can also monitor and evaluate the impact of human activities on the environments that surround us, especially on the roads to achieve the highest safety. As we can see from the Casualties dataset, there are 8 important features that are relevant and really could influence the prediction of new cases Casualty type could be: pedestrian, cyclist, taxi rider, horse rider, tram occupant, car occupant, .. etc. From this we can extrapolate that human type, could be a prominent factor that impacts and influences traffic flow or incidents, since each type has certain pattern. Moreover, 7 out of 8 most important predictors (attributes) in this dataset are human related attributes, this strongly supports our case that studying human behavior and impact from historical data would greatly assist us in predicting the future. An interesting finding from our analysis of this data, is the age band of the casualty. The analysis shows that age band of the casualty plays an important role in causing the traffic accidents. Results show that the second important attribute in causing the accidents is the age of casualty involved in the incident. Figure 3, presents that 80% of the casualties are not car passengers, which means 80% of the casualties are among the pedestrians on the road and sidewalks. This high percentage is potentially due to the unavailability of convenient sidewalks, unavailability of crosswalks, and invalidity of some of the road components such as traffic lights or signs. It could be noticed clearly from the same figure that the front seat passengers casualties percentage is higher than the rear seat. In this manner, using seat belt should be strictly monitored, also, seat belt awareness should be done constantly. The percentage is 59.2% male to 40.8% female casualties. Furthermore, fatal accidents happen to male drivers and passengers more than female. Figure 2 shows the distribution for fatal and slight casualties over the sex of the casualty. This piece of info can assist insurance companies in certain areas, depending on studying drivers' behavior to design different policies for male and female. The same thing

could be utilized by examining different factors, all of these depend on the content of the analyzed data.

FIGURE-1



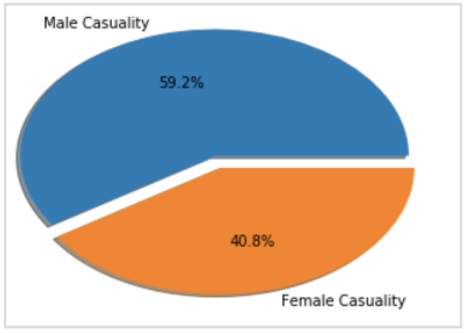


FIGURE-2

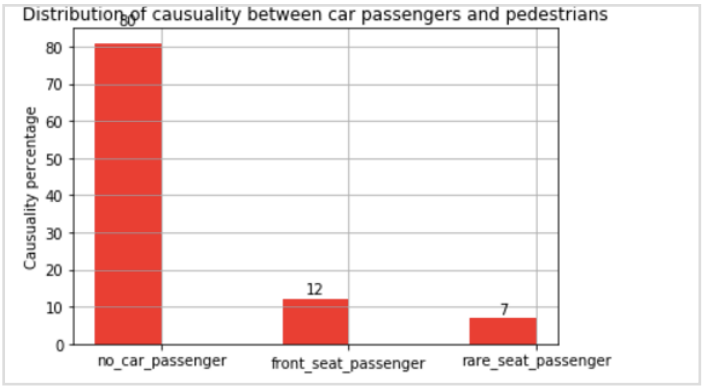


FIGURE-3

IV. CONCLUSIONS

Thousands of people die in traffic crashes yearly. People lose their lives every day and more people are injured every hour. In our study we use python and machine learning techniques to evaluate two classifiers on two big workbench datasets. The used classifiers are: Naive Baye's andRandom Forest .Our analysis and the extracted patterns and findings can assist decision makers and practitioners to enhance the transportation system intelligently and develop new rules. This study revealed some common misconceptions about road incidents. Our analysis showed that the human behavior has strong impact on the traffic flow and safety decisions. Our results revealed that driver's attributes such as age and sex could be predicted correctly up to 70% by providing other attributes for an accident or casualty.

REFERENCES

[1] U. DFT, UK Department For Transportation Traffic Dataset Repository, 2014. [Online]. Available: https://data.gov.uk/dataset/ road- accidents- safety-data

[2] E. Florido, O. Castano, A. Troncoso, and F. Martınez-Alvarez, "Data mining for predicting traffic congestion and its application to spanish data."

[3] H. Ibrahim and B. H. Far, "Data-oriented intelligent transportation systems," in Information Reuse and Integration (IRI), 2014 IEEE 15th International Conference on. IEEE, 2014, pp. 322–329.

[4] J. Alcala-Fdez, L. Sanchez, S. Garcia, M. J. del Jesus, S. Ventura, J. Garrell, J. Otero, C. Romero, J. Bacardit, V. M. Rivas *et al.*, "Keel: a software tool to assess evolutionary algorithms for data mining problems," *Soft Computing*, vol. 13, no. 3, pp. 307–318, 2009.

[5] S. Krishnaveni and M. Hemalatha, "A perspective analysis of traffic accident using data mining techniques," *International Journal of Computer Applications*, vol. 23, no. 7, pp. 40–48, 2011.

[6] M.Chong,A.Abraham, and M.Paprzycki,"Traffic accident data mining using machine learning paradigms," in Fourth International Conference on Intelligent Systems Design and Applications (ISDA'04), Hungary, ISBN, vol. 1047219710, 2004, pp. 415–420.

[7] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.

[8] scikit-learn: machine learning in Python — scikit-learn 0.18.2 documentation", *Scikit-learn.org*, 2017. [Online]. Available: http://scikit-learn.org/stable.

[9] Learn R, Python & Data Science Online | DataCamp", *Datacamp.com*, 2017. [Online]. Available: https:// www.datacamp.com/courses/intro-to-python-for-data-science.