Analytics of Road Accidents in USA using Visualization Techniques and Machine Learning Approach

M.R.G.Vijithasena
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
Rasika.20@cse.mrt.ac.lk

D.H.D. Kaveendri
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
Dilani.20@cse.mrt.ac.lk

Abstract- **Road accidents are major problems that lead to death, severe injuries and disabilities to human lives as well as physical properties all over the world. According to statistical information the United States of America ranked in a considerable position in Road Traffic accidents globally. This study discusses an in-depth analysis that identifies significant factors, causes behind these accidents and the quantification of factors that affect the severity of accidents based upon the data availability. Reviewing the various types of factors involved in different types of accidents and methodologies involved in analysis are considered. Analysis based in descriptive, predictive using machine learning and prescriptive for developing measures to improve their safety performance is targeted. This includes approaches to foresee the number of accidents that may happen in future and quantifying in what areas to change in order to reduce road accidents to a considerable amount.**

Keywords— Descriptive, Predictive, Prescriptive, Machine Learning

INTRODUCTION

Road accidents have become a major issue resulting in huge economic loss. In order to bring down the accident rate a detailed analysis of the factors which are responsible for the accidents are to be analyzed. There are a large number of attributes such as road conditions, traffic flow, environmental state and user's behavior on the road influencing Road accidents [1]. Considering those attributes, 'n' number of models are built, ignoring the complexity of the factors associated with causing accidents and involving only one or more variables. In modelling a traffic accident scenario, it is vital to analyze the complex situation with all the relevant attributes .It is identified that the factors like traffic volume, mix of modes, type of vehicles, pedestrians, traffic segregation measures and road geometries influence the accident scene. Therefore, any model to study the traffic accident scene should be comprehensive and accommodate all these identified factors. Moreover, the following three factors have been found to be very crucial in ensuring road safety and reducing accidents: Traffic Engineering, Traffic Education and Traffic Enforcement [3]. These analyses are succeeded by the predictive analytics mechanism to speculate the number of accidents in the upcoming years. Proper analysis of the road accident data provides a useful insight regarding the causes and consequences of accidents. It would be beneficial to accurately predict the severity of accidents which would lead to encountering the issues in advance. Predictive analytics is the latest field which can be applied to this accident data to predict the future accidents. Machine learning is the artificial intelligence technique that is used to create a model which learns from the past data. Once the model is created whenever a new set of data is given to it, it would be able to predict the approximate values. Regression, classification, clustering, recommender system, and churn prediction are among the machine learning algorithms which can be used. If these artificial intelligence and machine learning concepts are applied on the data, a more accurate forecasting can be made. The overall outline of this paper can be organized as follows; section two highlights the description about the data used. Section Three discusses the various accident data analysis techniques and the results using easily understandable visual representations. Section four discusses the predictive

analysis and its accuracy levels and then section five highlights prescriptive analytics techniques. Finally, the conclusion is explained.

DATA COLLECTION

A. *Study Conducted Area*

This is a countrywide car accident dataset, which covers 49 states of the United States. The accident data are collected from February 2016 to December 2019, using several data providers, including two APIs that provide streaming traffic incident data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.0 million accident records and 49 features in this dataset.

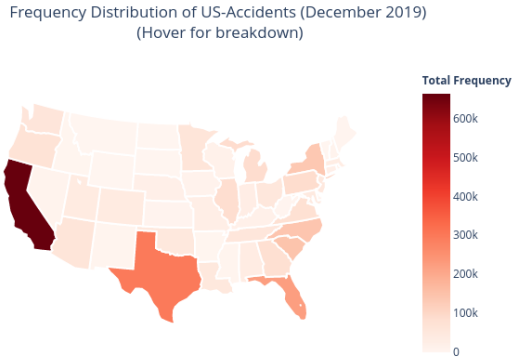

Fig.1. Frequency Distribution of US Accidents

B. *Features in dataset*

Weather conditions : Temperature(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Speed(mph), Weather condition, Weather timestamp, Sunrise_sunset, Wind direct, Wind_Chill(F)

Infrastructure factors : Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop

Location and time : Start_Time, End_Time, Start_Lat, Start_Lng, End_Lat, End_Lng, Distance, Number, Street, Side, City, Country, State, Zip code, Timezone, Airport_code, Astronomical_Twilight

C. *Data cleaning*

1. Handling missing values

   Some columns have contained a huge amount of null values when comparing with the total number of rows. So in the very beginning, those columns were identified and removed. Ex: End_Lat, End_Lng, Precipitation (in), Number, Wind_Chill (F)

   Other rows which contained missing values have been removed.

2. Handling incorrect values

   Some rows have contained incorrect values such as getting negative values for the gap between start and end times of accident. So those rows have been removed before the analysis stage.

ROAD ACCIDENT ANALYSIS

The road accident analysis was performed in three stages; the first stage is all about the detailed analysis of the contributing factors of accidents, nature, and class, the second stage is about the trend analysis related to road accidents followed by a prediction of road accident analysis. The tools used for the visual analysis and prediction using Python.

A. *Severity Distribution*

Severity "a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).Distribution of the severity values can be seen in fig,2. Severity 2 type accidents count up to more than 12000 which have the highest frequency and lowest shown by severity type 1 accidents making the severity 3 accidents to the second highest frequency. Severity 4 which has the highest risk has low count below 5000 and medium risky severity are of highest frequency.

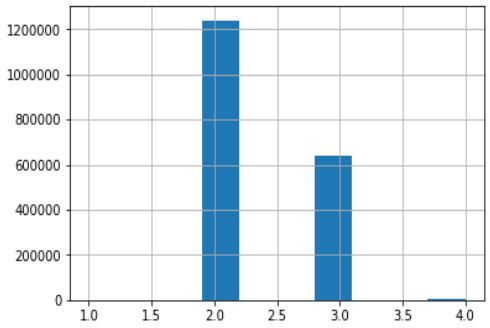

Fig.2: Severity Distribution

```
2    1239576
3     642202
4       4396
1        804
Name: Severity, dtype: int64
```

B. *Analysis based on State*

According to Fig.3 CA is the state which has the highest number of accidents above 80,000 and Texas is the second highest number of accident reported state.

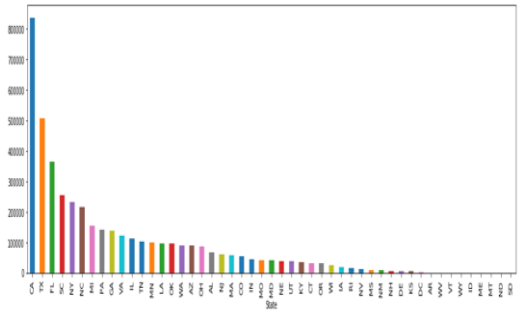

Fig.3: State Wise count on accidents

C. *Severity wise Analysis based on day of the week*

From the analysis below in Fig.4 it indicates the first five days of the week has four times higher accidents count than the other two days, in each severity type. Weekend shows low risk in happening accidents rather than weekdays. Severity wise the charts are merely similar with small variations.

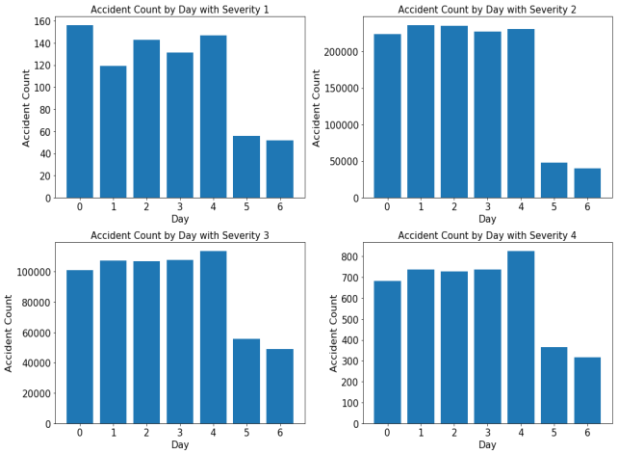

Fig.4: Severity wise Number of Accidents distribution

D. *Severity distribution based on weather conditions*

Analyzing whether a severity distribution in accidents are very crucial in order to have a clear picture about casualties of accidents. Fig.5 indicates in clear weather most of severity 2 and 3 are prominent and least in cloudy weather conditions. Weather conditions have a less impact on severity 1 and severity 4.

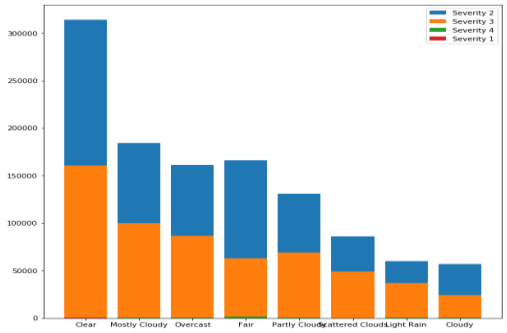

Fig.5: Weather based severity distribution

E. *Mean Severity Vs. Temperature, Humidity and Pressure*

Climatic factors like Temperature, Humidity and Pressure which were drawn as a line graph as a function of mean severity is shown in Fig.6.Severity increases as a function of Humidity with fluctuations. Pressure and Temperature are decreasing functions of mean severity.

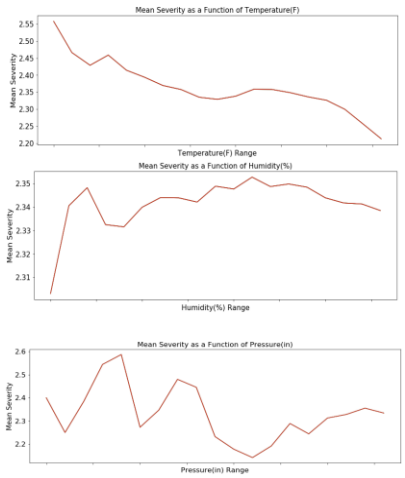

Fig.6: Mean Severity as a function of Pressure, Temperature, Humidity

F. *Severity wise accident analysis based on Infrastructure type*

According to the graphs, severity 2 shows high frequency in other three types of infrastructure except for near junctions which indicates highest frequency in severity 3.Values show high number of severity 2 accidents near Traffic signals and low number of count near rounder bound. Analysis implies highest risk areas near traffic signals and junctions.

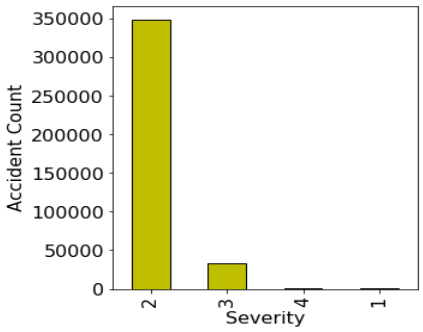

Fig. 7: Near Traffic signal

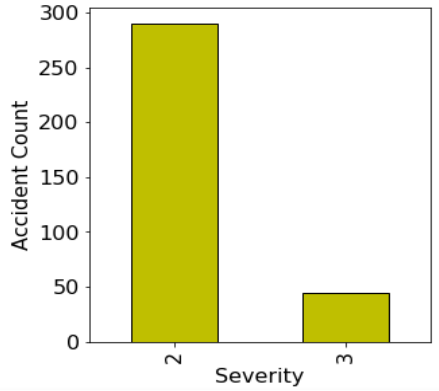

Fig. 8: Near Bump

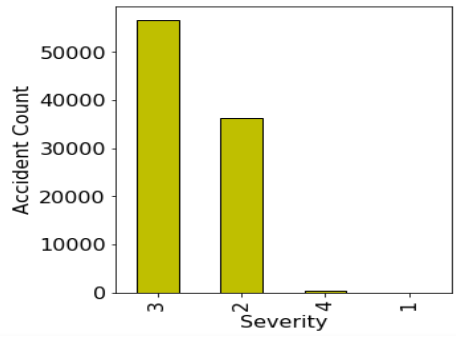

Fig. 9: Near Junction

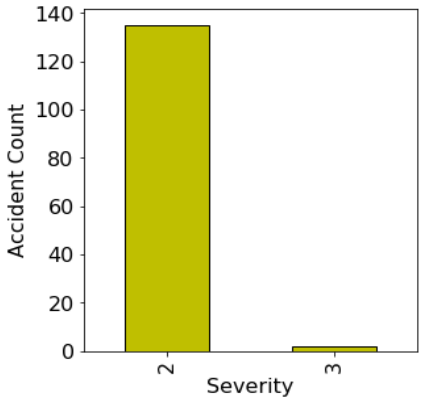

Fig. 10: Near Rounder bound

G. *Correlation coefficient of the variables*

The correlation of the target variable with the given set of variables are low overall. There are some variables with no correlation like Humidity and Time duration implies those factors have no impact on severity count. Highest negative correlation among all factors, -0.268792 is represented by Traffic

signal indicates inversely proportional low impact where Severity decreases when Traffic Signal increases and positive correlation, 0.181313 is shown by TMC where TMC increases the severity will also increases

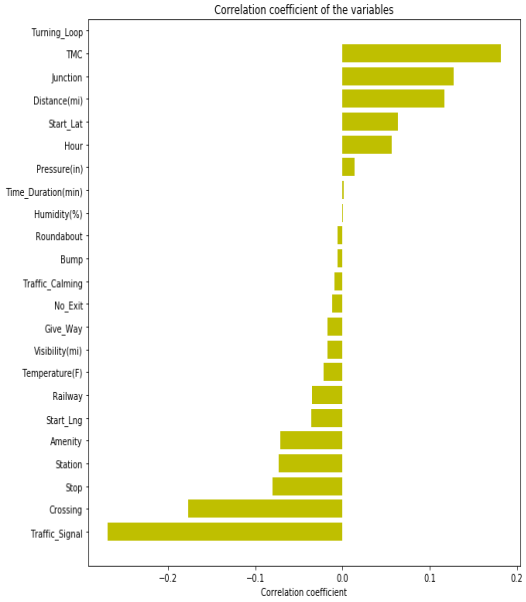

Fig.11: Correlation coefficient of variables

V. PREDICTIVE ANALYTICS USING MACHINE LEARNING

There are three main approaches to data analytics, i.e. descriptive analytics, predictive analytics and prescriptive analytics. Descriptive analytics involves analyzing the past to examine what has happened and provide insights on how to approach the future. The main objective of descriptive analytics is to reveal the reasons behind the success or failure occurred in the past. Predictive analysis analyses the past data patterns and trends and provides future speculations using them whereas prescriptive analytics is the next stage of predictive analytics that integrates the ability to manipulate the future. Predictive analysis uses statistics, machine learning, data mining, etc. in predicting future trends and outcomes. Machine learning techniques have become popular in applying predictive analytics due to the exceptional usage in large scale. Therefore, machine learning and artificial intelligence algorithms can be used to optimize and find new statistical patterns in the road traffic accident analysis.

Since the dataset contains around 3 million records and a limited CPU in the laptop, a selected dataset has been used for further prediction tasks. (State: Pennsylvania, country: Montgomery)

Model Selection

There are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning approaches available for creating a machine learning model. By considering the input data and the type of learning required, supervised learning methodologies have been selected. Since the output variable takes continuous values, prediction models have been used to predict severity of given input features. Machine learning algorithms: Logistic Regression, KNN, Decision tree, Random Forest

A. Accuracy Comparison

In the following chart, accuracy values of mentioned models have been compared.

TABLE 1. Accuracy model table

Model	Accuracy
Log Regression	0.955
KNN	0.938
Decision tree	0.962
Random Forest	0.972

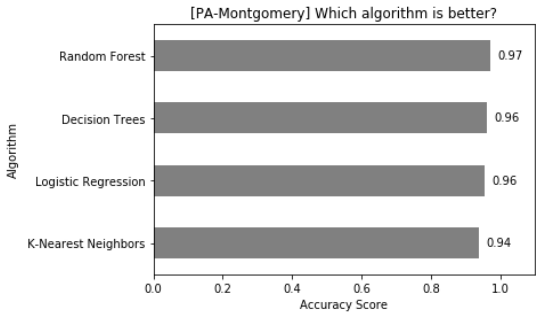

Fig.12: Accuracy chart

According to the accuracy results, the model which was created by using the random forest method provided the highest accuracy (97%).

TABLE 2. Precision, recall, F1-score

Model	Severity	precision	Recall	f1-score
Logistic Regression	1	0.0	0.0	0.0
	2	0.97	0.99	0.97
	3	0.75	0.56	0.64
	4	0.0	0.0	0.0
KNN	1	0.0	0.0	0.0
	2	0.94	1.00	0.97
	3	0.78	0.19	0.31
	4	0.0	0.0	0.0
Decision tree	1	0.0	0.0	0.0
	2	0.97	0.98	0.98
	3	0.76	0.67	0.71
	4	1.00	0.33	0.5
Random forest	1	0.0	0.0	0.0
	2	0.98	0.99	0.98
	3	0.88	0.68	0.77
	4	0.0	0.0	0.0

As F score is the harmonic mean of precision and recall, it is considered as best measure of test accuracy for severity 2 and 3 scenarios F score is high in Random forest and it implies as the best test for predicting severity of accidents.

PRESCRIPTIVE ANALYSIS

An emerging discipline and represents a more advanced use of predictive analytics. Prescriptive analytics goes beyond simply predicting options in the predictive model and actually suggests a range of prescribed actions and the potential outcomes of each action. Prescriptive model is able to predict the possible consequences based on a different choice of action, it can also recommend the best course of action for any pre-specified outcome. Google's self-driving car, Waymo, can be proposed as prescriptive tool where this vehicle makes millions of calculations on every trip that helps the car decide when and where to turn, whether to slow down or speed up and when to change lanes — the same decisions a human driver makes behind the wheel [15].

CONCLUSION

Roads are the main means of transport in the USA which in turn leads to substantial number accidents. This paper analyzed almost three million data with indication in various road accident type and casualty severity. The key findings emerged from our study indicates that the mediate severity of accidents have a high frequency of occurring rather than severities of very low and high risks and factors like weather, infrastructure, day of week and other climatic factors have various effects on these accidents .Random Forest algorithm has an accuracy of 0.972 which is the highest accuracy in predicting these road traffic accidents in USA and action to reduce number of accidents, use of self-driving cars like Waymo to take decisions as a human which identifies best cause of actions when needed.

REFERENCES

[1] M. Peden, R. Scurfield, D.Sleet, D. Mohan, A.A Hyder and E. Jarawan et al. "World report on road traffic injury prevention", World Health Organization, Geneva, Switzerland, 2004.

[2] Sultanate of Oman, Statistical Year Book, National Center For Statistics & Information, Issue 44, August 2016.

[3] P.Kai, "Human Factors for Road traffic Accidents in the Sultanate of Oman under consideration of Road Construction Designs",Ph.D thesis, Inst. Of Psychology, University of Regensburg, Bavaria, Germany 8 May 2014.

[4] J.Almatawah, "Towards Improving Crash Data Management System in Gulf Countries", Int. Journal of Engineering Research and Applications, Vol. 4, Issue 9(Version 5), pp.35-40, September 2014.

[5] M.M. Islam and A.Y.S. Hadhrami, "Increased Motorization and Road Traffic Accidents in Oman", Journal of Emerging Trends in Economics and Management Sciences (JETEMS),Vol.3, Issue. 6, pp.907-914,2012.

[6] S.G. Farag, I.H. Hashim, S.A. El-Hamrawy, "Analysis and Assessment of Accident Characteristics: Case Study of Dhofar Governorate, Sultanate of Oman", International Journal of Traffic and Transportation Engineering,Vol.3, Issue. 4, pp.189-198,2014.

[7] A. Galal, 2010. “Traffic Accidents and Road Safety Management: A Comparative Analysis And Evaluation in Industrial, Developing and Rich-Developing Countries” in 29th South African Transport Conference (SATC 2010) ISBN: 978-1-920017-47-7, 16-19 August 2010, Pretoria, South Africa, 2010.

[8] E. Kopits, M. Cropper, “Traffic fatalities and economic growth, Accident Analysis and Prevention”, Eastern Mediterranean Health Journal, vol.37, issue 1, pp.169–178, January 2005.

[9] Bener, A., Hussain, S.J., Al-Malki, M., Shotar, M.M., Al-Said, M.F., Jadaan, K.S., “Road traffic fatalities in Qatar, Jordan and the UAE: estimates using regression analysis and the relationship with economic growth”. Eastern Mediterranean Health Journal., Vol.16, no.3, pp.318– 323, 2010.

[10] R. Elvik, P. Christensen, A. Amundsen, “Speed and Road Accidents: An Evaluation of the Power Model”. Institute of Transport Economics, Oslo, 2004

[11] k.Rajimon,’ Oman traffic: OMR500, one year jail term for drivers jumping red light’,Times of Oman, 2016.

[12] Available: https://www.dezyre.com/article/types-of-analytics- descriptive-predictive-prescriptive-analytics/209, retrieved on May 18, 2017

[13] Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-656.2008.01390.x/full, retrieved on May 18, 2017 [16] Available: http://www.stat.cmu.edu/~hseltman/618/LNTS4.pdf, retrieved on May 18,

[14]"Difference of Predictive vs. Prescriptive Analytics | Ohio University", *Ohio University*. [Online]. Available: https://onlinemasters.ohio.edu/blog/predictive-vs-prescriptive-analytics-whats-the-difference/. [Accessed: 05- Apr- 2020].

[15]"Home – Waymo", *Waymo*. [Online]. Available: https://waymo.com/. [Accessed: 05- Apr- 2020].