Accident Severity in India

Yash Kumar Arora, Santosh Kumar, Umesh Kumar Tiwari, Shubhank Singhal , Vijay Kumar

Abstract*: Among all the transportation services available, road transport is one of the most important services available. It acts as a feeder for the other services. According to the Ministry of Road Transport and Highways, the road transport amounts to the traffic of about 87% related to the passenger and 60% related to the freight. Now, there is another field, where road transport is among the top list and that field is of road accidents. In 2016, about 150785 people died in road accidents. And as the population is increasing, there is also an increase in the rate of road accidents. So, it is vital to analyze the data of road accidents for future predictions and thereby developing proper measures for this increasing rate. Many factorsresult in accidents and many cases might not have been recorded. So, the available data may not be consistent, but the data is gathered mainly from the Ministry of Road Transport and Highways and then the information was extracted from that data. This information is used for the statistical analysis for the prediction of a future road accident or accident severity.

***Index Terms*: Polynomial Regression, R2 Score,Road Accidents**

I. INTRODUCTION

Road accidents almost take place every day in some parts of the country. There may not be any day without any road accident. Some of them may be harmless but most of them prove to be fatal. There isn number of reasons for any road accident to take place. Some of them can be the recklessness of the driver or the faulty transportation system. According to the stats, there were a total of 4,67,044 accidents in the year 2018 in which, 1,51,417 people died. This shows the severity of road accidents. With the increase in population, road accidents are still increasing every year. In some cases, an accident might not be reported or properly recorded by the police. And there is no robust system which monitors the regular collection of reliable and systematic data. This led to the inconsistent data, but available data will be less than the original one, which clearly shows the seriousness of the matter.

However, for proper measures against road accidents, studies and research should be driven. One of the major problems with these studies is access to accurate information. In fact, except for the available online statistics, access to most of the information requires a heavy administrative procedure.

Revised Manuscript Received on October 20, 2019.

Yash Kumar Arora, Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India.

Santosh Kumar, Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India.

Umesh Kumar Tiwari, Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India.

Shubhank Singhal, Computer Science and Engineering, Graphic Era deemed to be University, Dehradun, India.

Vijay Kumar, Department of Physics, Graphic Era Hill University, Dehardun

II. RELATED WORK

Abdalla et al. [1] reveal that traffic accidents are more frequent near residential areas in comparison to areas that are not near residential areas. Mussone et al. [2] analyzed road accidents that occurred at intersections in Milan, Italy region. They used deep learning to analyze the accident data. Their results showed that the highest frequency of accidents in that region was of the pedestrian hit accident at a non-signalized intersection and nighttime. Yash et al. [3] used linear regression to predict road accidents in India. Pochet al. [4] states traffic accidents at urban intersections result in a huge cost to society in terms of death, an injury so it's fatal for human life. Many agencies are working on methodologies to reduce the effect of accidents. In this paper, seven-year historical data have been used to predict the effectiveness of specific intersection, but it didn't go on showing the Geometric traffic-related elements, however, figures do not include a person's desire to live.

MiaouSp [5] research deals with the relationship between truck accidents and geometric design of road sections, then evaluate the performance of negative binomial regression which in turn is used to estimate the unknown parameter. NB regression model should be used with high caution. So, at the initial stage while examining the data Poisson Regression model should be used and if the results give moderate or high accident data then the NB model can be handy.

Karlaftis et al. [6] talks about the heterogeneity issue coming in panel data sets due to which it is being unimportant for analysis of safety research and there have been no previous research or study on it. This issue needs to be solved, otherwise, it would result in incorrect statistical data. So, this paper talks about the issue of heterogeneity and how to resolve it by following some steps including homogenous clusters are formed for different observations and applying a binomial group model to each cluster and then calculating differences between them.

Jones B, Janssen L and Mannering F [7] collective research shows the enormous impact on urban freeways and talks about how the accident frequency on these freeways can be used to develop or seek comprehensive strategies for getting ideologies for reducing the accident rate on these freeways. Kockelman et al. [8] research indicates traffic accidents remain a major problem for people. Talking about stats around 2 million died on U.S roads, 3 million were reported for traffic accidents. So, the research deals with crash frequency and crashes severity to seek methods to reduce the accident rate.

Homogenous high-speed

road data was collected and was clustered using cluster analysis technique and a crash severity model was estimated for total crash counts into counts by severity. The results displayed how increasing speed limit over a particular area raise the accident rate like for areas that are designed for a limit of 55 miles per hour, if the speed increases more to 10 miles per hour, 3.29% accident rates on roads take a hike. But there are certain drawbacks also for using cost evaluation models as they result in inconsistency estimation due to inappropriate assumptions.

Chen et al.[9] talks about the advancement of vehicle technologies which would help a driver to avoid crashes plus for comfort more entertaining applications can be installed for drivers inside a car. Considering, the fact these applications and new reforms would make the driver concentrate to focus at a single point thus crashes and accidents can be reduced and avoided. Several other studies focused on accident severity analysis using traditional statistical techniques and provide good results [10-13].

In this paper, a polynomial regression algorithm is used to predict accident severity which could increase in the future. We split the dataset in training and testing in the ratio 80:20. Then, we fit the curve and find the intercept and slope of the curve to predict the accident severity. Hence, our main emphasis will be on the prediction of accident severity.

III. METHODOLOGY

A. Polynomial Regression

It is a type of linear regression in which we fit a polynomial equation on the data with a curvilinear relationship between the dependent variable and the independent variables.

The polynomial equation of degree n represented as in eq.1.

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots\ldots + \theta_n x_n \quad (1)$$

where, θ_0 is the bias, θ_1, θ_2, …, θ_n are the weights in the equation of the polynomial regression, and n is the degree of the polynomial.

B. R2 Score

R-squared is a statistical measure of how close the data are to the fitted regression line.

$$R\ squared = \frac{Explained\ variation}{Total\ variation}$$

R-squared rangesfrom 0 to 1.

0 indicates that the model explains none of the variability of the response data around its mean. 1 indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits your data and if it is negative, then the model is completely wrong.

R-squared is defined in eq. 2.

$$R^2 = 1 - \frac{SS_r}{SS_t} \quad (2)$$

where,

SS_t is the total sum of squares,

SS_r is the total sum of squares of residuals.

C. Algorithm

At first, for a good analysis, Data must be collected. After collection, the data must be processed to extract the information out of it. For this paper, we have used python as the programming language for our analysis. First, import the data. Then the data is split into two parts in the ratio of 80:20. One part is termed as train, as it will be used to train our model. And the other part is termed as a test, as it will be used for validation purposes. After splitting, the rest of the work will be done on the training dataset. Using python libraries and functions, fill the null values, and remove the outliers. The unrelated columns are dropped, and the transformation is applied to make the data of relatable scale. After cleaning, graph and plots can be drawn to have a visual representation of data. Then, the machine learning model will be made.

The polynomial regression model has been used in this paper. The training dataset is split into two parts. One part has variables that will be used for predictions and another part has the target variable which is to be predicted. After training the model, its accuracy must be checked. And for this, the test dataset will be used. Now, fill the null values, if present in the test dataset, using the same values as used in the training dataset. Outliers are not looked into in the test dataset. After filling of null values, the test dataset is split into two parts as the training dataset. The first part will be used in our model and predictions will be generated. Then, these predictions and the true values of the test dataset are used to check the accuracy. Here, accuracy is measured using R2score. If the value of the R2score is not good, then we must repeat from the cleaning and processing of the train data again. And if the value is good, then this model can be used to predict the number of accidents for a given year.

Step1	Input data
Step2	Split data into train and test. (train=80%, test=20%)
Step3	Cleaning of data. (i.e. removing outliers, filling Null values)
Step 4	Plot a graph between Year (X-axis) and the number of people who died in that year (Y-axis)
Step 5	Creating a machine learning model from the training dataset
Step 6	Validating the machine learning model with the test dataset
Step 7	Calculate the R2 Score
Step 8	If R2 Score is not good, repeat steps from step3, otherwise go to step9
Step 9	Predict the number of road accidents for a specified year from the above machine learning model

IV. DATASET

Dataset is collected from a government site of the Ministry of Road Transport and Highway (MORTH) [14]. Data contains the year-wise stats of the total number of accidents, the total number of fatal accidents, the number of people died, the number of people injured, and the accident severity. Fig. 1 is the small portion of the dataset used in this paper.

	year	total_no_of_accidents	fatal_no_of_accidents	no_of_person_killed	no_of_person_injured	accident_severity
0	2001	405637	71219	80888	405216	19.9
1	2002	407497	73650	84674	408711	20.8
2	2003	406726	73589	85998	435122	21.1
3	2004	429910	79357	92618	464521	21.5
4	2005	439255	83491	94968	465282	21.6

Fig. 1 Dataset

V. RESULTANALYSISANDFUTUREPREDICTION

Fig. 2 displays the plot between the total number of accidents or persons injured and the corresponding year. Fig. 3 displays the total fatal accident or person died and the corresponding year. Fig. 4 shows the accident severity year wise. These all are the graphs of the dataset used in this paper. Since the dataset is small, the machine learning model made from this dataset gives an R2 Score of 0.98, which is a good result and shows that this model can be used for prediction. Using this model, it is seen that if proper measures are not taken in time, then the total number of people who died in the year 2025 will be 59853, which is not an ignorable value.

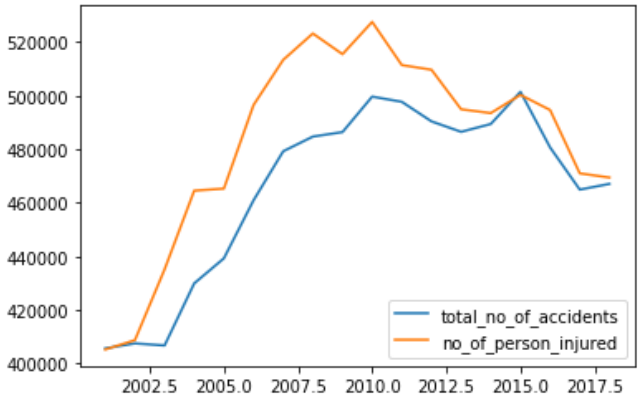

Fig. 2Total number of accidents and person injured year wise in India.

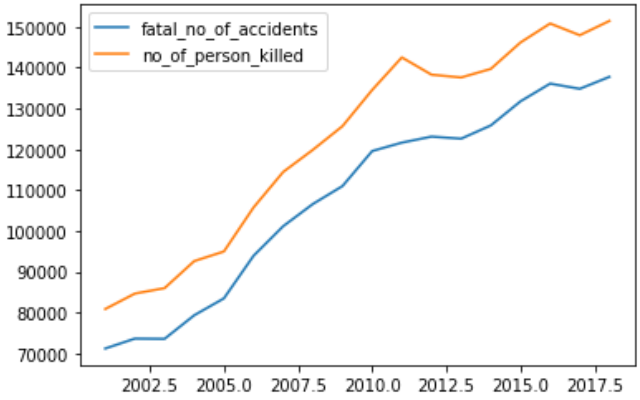

Fig. 3Total fatal accidents and person died year wise in India.

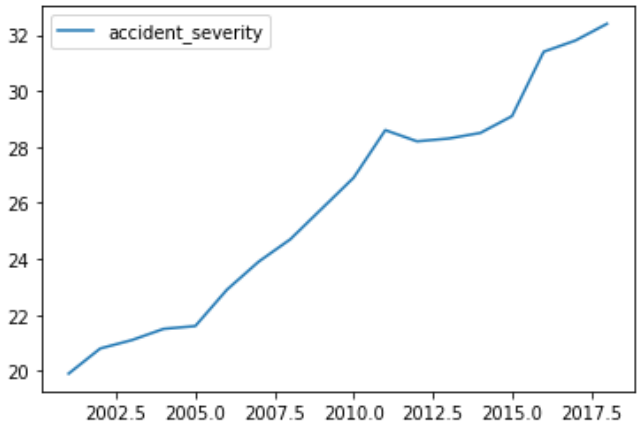

Fig. 4Accident Severity year wise in India

VI. CONCLUSION

In this paper, it is observed that if the government does not take proper measures, then there will remain a great number of cases every year with the death of a lot of people. As of 2017, the percentage of accident severity is 31.8 and the number of cases was464910 in which 147913 people died.

And since the graph of accident severity is of increasing nature, so if the government does not take safety measures then according to this paper, the accident severity will increase to 33.24 in 2020 and to 35.27 in 2025 which cannot be neglected. Therefore, the government should focus on road conditions and factors affecting road accidents in India otherwise the road accidents will increase at an alarming rate.

REFERENCES

1. Abdalla IM, Raeside R, Barker D, McGuigan DR (1997) An investigation into the relationships between area social characteristics and road accident casualties. Accid Anal Prev 29:583–593
2. Mussone L, Ferrari A, Oneta M (1991) An analysis of urban collisions using an artificial intelligence model. Accid Anal Prev 31:705–718
3. Arora, Yash & Kumar, Santosh. (2020). Statistical Approach to Predict Road Accidents in India. 10.1007/978-981-32-9515-5_18
4. Poch M and Mannering F (1996) Negative binomial analysis of intersection-accident frequencies. J TranspEng 122
5. Miaou SP (1994) The relationship between truck accidents and geometric design of road sections–poisson versus negative binomial regressions. Accid Anal Prev 26
6. Karlaftis M, Tarko A (1998) Heterogeneity considerations in accident modeling. Accid Anal Prev 30:425–433
7. Jones B, Janssen L and Mannering F (1991) Analysis of the frequency and duration of freeway accidents in Seattle. Accid Anal Prev 23
8. J. Ma, K. Kockelman (2006) Crash frequency and severity modeling using clustered data from Washington state. In: IEEE Intelligent Transportation Systems Conference. Toronto Canadá
9. Chen W, Jovanis P (2002) Method of identifying factors contributing to driver-injury severity in traffic crashes. Transp Res Rec. 1717
10. Abdel-Aty MA and Radwan AE (2000) Modeling traffic accident occurrence and involvement. Accid Anal Prev 32
11. Maher MJ and Summersgill IA (1996) Comprehensive methodology for the fitting of predictive accident models. Accid Anal Prev 28
12. Joshua SC and Garber NJ (1990) Estimating truck accident rate and involvements using linear and poisson regression models. Transp Plan Technol 15
13. Miaou SP and Lum H (1993) Modeling vehicle accidents and highway geometric design relationships, Accid Anal Prev 25
14. MORTH Road Accidents in India 2016. New Delhi: Ministry of Road Transport and Highways, Transport Research Wing, Government of India, August 2018.

AUTHORS PROFILE

Yash Kumar AroraI am pursuing BTech in Computer science with specialisation in Big Data & Analytics from Graphic Era deemed to be University, Dehradun. With, interest in research, I want to research more and more to gain knowledge that can be used efficiently. Currently, I am an intern at IBM India Software Labs, Bengaluru.

Dr. Santosh Kumar had received his Ph.D. from IIT Roorkee (India) in 2012, M. Tech. (CSE) from Aligarh Muslim University, Aligarh (India)in 2007 and B.E. (IT) from C.C.S. University, Meerut (India) in 2003. He has more than 13 years of experience in teaching/research of UG (B. Tech.) and PG (M.Tech.) level courses as a Lecturer/Assistant Professor/ Associate Professor in various academic /research organizations. He has supervised 01 Ph.D. Thesis, 20 M.Tech. Thesis, 18 B.Tech projects and presently mentoring 06 Ph.D students, 03 M.Tech students and 04 B.Tech. students. He has also completed a consultancy project titled "MANET Architecture Design for Tactical Radios" of DRDO, Dehradun in between 2009-2011.He is an active reviewer board member in various national/International Journals and Conferences. He has memberships of ACM (Senior Member), IEEE, IAENG, ACEEE, ISOC (USA) and contributed more than 46 research papers in National and International Journals/conferences in the field of Wireless Communication Networks, Mobile Computing and Grid Computing and software Engineering. Currently holding position of Associate professor in the Graphic Era Deemed to be University, Dehradun (India). His research interest includes Wireless Networks, MANET, WSN, IoT, and Software Engineering.

Dr. Umesh Kumar Tiwari is working as an Associate Professor in Department of Computer Science and Engineering in Graphic Era Deemed to be University, Dehradun. He had received his Ph.D. in 2016. He has more than 12 years of experience in teaching/research of UG and PG level degree courses as a Lecturer/Assistant Professor/ Associate Professor in various academic/research organizations. He is supervisor of 02 PhD and 5 M. tech students who are working on specific domains of software engineering and network security. His research is on multidisciplinary topics and he has published 17 journal papers in reputable international and national journals, and 12 conference papers in reputed international and national conferences. His research interests are Wireless Communication Networks, Network Security, and Software Engineering topics with improved modeling, interaction-integration complexities, testing and reliability models.

Shubhank SinghalI am pursuing BTech - Computer science and Engineering from Graphic Era Deemed to be University, Dehradun.