

Road Accident Prediction using Machine Learning Algorithm

Vipul Rana¹, Hemant Joshi², Deepak Parmar³, Pradnya Jadhav⁴, Monika Kanojiya⁵

^{1,2,3,4}UG Scholar, Computer Engineering Department, Shah and Anchor Kutchhi Engineering College, Mumbai, Mumbai University, Maharashtra, India

⁵Professor, Computer Engineering, Shah and Anchor Kutchhi Engineering College, Mumbai, Mumbai University, Maharashtra, India

Abstract - The traffic has been transformed into the difficult structure in points of designing and managing by the reason of increasing number of vehicles. This situation has discovered road accidents problem, influenced public health and country economy and done the studies on solution of the problem. Large calibrated data agglomerations have increased by the reasons of the technological improvements and data storage with low cost. Arising the need of accession to information from this large calibrated data obtained the corner stone of the data mining. In this study, assignment of the most compatible machine learning classification techniques for road accidents estimation by data mining has been intended.

experience. Machine learning is able to attain extract information from data and use statistical method.

1.1 Objectives

The main objective of the road accident prediction system:

- Analyze the previously occurred accidents in the locality which will help us to determine the most accident-prone area and help us to set up the immediate required help for them.
- To make predictions based on constraints like weather, pollution, road structure, etc.

1.2 Problem Statement

There are several problems with current practices for prevention of the accidents occurred in the localities. The database we will use is available officially by many institutes and government websites. The data collected will be analysed, integrated and grouped together based on different constraints using the best suited algorithm. This estimation will be helpful to analyse and identify the flaw and the reasons of the accidents. It will also be helpful while making roads and bridges as a reference to avoid the same problems faced before. The predictions made will be very much useful to plan the management of such problems.

2. Logistic Regression

Logistic regression is the regression analysis and dependent upon the variables is binary numbers i.e. (0s and 1s), All regression analysis, the logistic regression is a prediction analysis. Logistic regression is used to details about data and to graphically explain the relationship between dependent binary variable and more nominal, ordinal, interval independent variables.

Sometimes logistic regressions are difficult to describe the statistics tools are easily conduct and analysis the datasets, then in others plain word are as it is display in the output.

1. INTRODUCTION

One of the most complicated and difficult daily needs is overland transportation.

In India, more than 150,000 people are killed each year in traffic accidents. That's about 400 fatalities a day and far higher than developed auto markets like the US, which in 2016 logged about 40,000. Every year over 1 million vehicles are added to traffic averagely. 1.2 million People have died and over 50 million people have been injured in road accidents in the world every year. Studies on traffic have executed that road accidents and death-laceration ratio will increase.

Design and control of traffic by advanced systems come in view as the important need. Assumption on the risks in traffic and the regulations and interventions in the end of these assumptions will reduce the road accidents. An assumption system which will be prepared with available data and new risks will be advantageous.

Data mining concept had been come up with by increasing and storage of data in the digital stage. Data mining involves the studies which will discover information from systematic and purposeful data structures obtained from disordered and meaningless data.

Machine learning which is sub-branch of artificial intelligence supplies learning of computer taking advantage of data warehouses. Assumption abilities of computer systems have advanced in the event of machine learning. Utilization of machine learning is a widespread and functional method for taking authentic decisions by using

$$h\theta(x) = \text{sigmoid}(Z)$$

Sigmoid Function

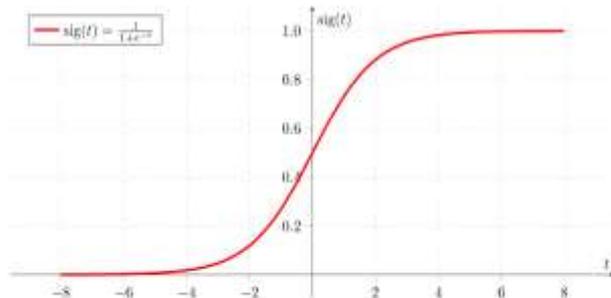


Figure 2: Sigmoid Activation Function

3. Software and Languages used

3.1 Jupyter

Jupyter exists to develop open-source software. It is used for open-standards, and services for interactive computing across dozens of programming languages. It is an open-source web application that allows you to create and share documents and code live. Which is a very big advantage of Jupyter. It can be used for data cleaning and transformation, numerical simulation, statistical modeling, machine learning and much more. We used Jupyter to run the algorithm.

3.2 Python

Python is an interpreted, high-level and a general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It is the most used programming language presently. It provides constructs that enable clear programming on both small and large scales. The logistic regression in the system is implemented in jupyter and the algorithm is written in python language.

3.3 HTML,CSS,JSCRIPT

These are the most commonly used webpage development languages. The user interface designed for the prediction system is been developed using these languages. The website acts as an interface which takes the input those are the various constraints from the users and passes to the program to work upon.

4. WORKING

The working of the project is divided in to parts.

4.1 Data Set Selection

Data is the most import part when you work on prediction systems. It plays a very vital role your whole project i.e., you system depends on that data. So selection of data is the first and the critical step which should be performed properly, For our project we got the data from the government

website. These datasets were available for all. There are other tons of websites who provide such data. The dataset we choose wad selected based on the various factors and constraints we were going to take under the consideration for our prediction system.

4.2 Data Cleaning and Data Transformation

After we have selected the dataset. The next step is to clean the data and transform it into the desired format as it is possible the dataset we use may be of different format. It is also possible that we may use multiple datasets from different sources which may be in different file formats. So to use them we need to convert them into the format we want to or the type that type prediction system supports. The reason behind this step is that it is possible that the data set contains the constraints which are not needed by the prediction system and including them makes the system complicated and may extend the processing time. Another reason behind data cleaning is the dataset may contain null value and garbage values too. So the solution to this issue is when the data is transformed the garbage values are replaced. There are many methods to perform that.

4.3 Data Processing and Algorithm Implementation

After the data is been cleaned and transformed it's ready to process further. After the data has been cleaned and we have taken the required constraints. We divide the whole dataset int o the two parts that can be either 70-30 or 80-20. The larger portion of the data is for the processing. The algorithm is applied on that part of data. Which helps the algorithm to learn on its own and make prediction for the future data or the unknown data. The algorithm is executed in which we take only the required constraints from the cleaned data. The output of the algorithm is in 'yes' and 'no'. It gives the error rate and the success rate.

4.4 Output and User Side Experience

After the prediction system is ready to use. The Website is developed for the user. The user just has to fill a form which consists of different options they need to select. They are like the type of climate, the type of vehicle and so on. Once the user submits the form the algorithm is triggered and the input given by the user is passed to the prediction system. The user is given how accident prone the road can be in percentage.

5. RESULTS

In the road accident prediction project use the dataset is in terms of values and some data is plain English word so, the numerical values data is easily predicted and also the calculation are easily done but, the normal word are display as it is or the non predicted data are drop in the table.

So, This dataset are many columns and rows and all numbers of null values will be fulfill in forward fill method and also use the classification algorithm entire dataset.

In that classification algorithm we will use Logistic Regression Algorithm The logistic algorithm will make the prediction in terms of percentage, to find accuracy level in percentage and Error percentages. This Algorithm is only for the yes and no type of result or successful and unsuccessful. The equation for combinations of all 15 input variables.

The classification algorithm of the entire dataset. In the Road Accident prediction final result is to find the percentage of accident in particular area. Having lower number of features helps the algorithm to converge faster and increases accuracy.

In the Road Accident prediction final result is to find the percentage of accident in particular area. Then we apply logistic regression on these features and obtain the least error.

6. CONCLUSION

Road Accidents are caused by various factors. By going through all the research papers it can be concluded that Road Accident cases are hugely affected by the factors such as types of vehicles, age of the driver, age of the vehicle, weather condition, road structure and so on. Thus we have build an application which gives efficient prediction of road accidents based on the above mentioned factors.

REFERENCES

- [1] Peden, M. (2004) "World report on road traffic injury prevention". Geneva: World Health Organization
- [2] M. Chang, L. Y., & Chen, W. C. (2005). "Data mining of tree- based models to analyse freeway accident frequency". Journal of Safety Research, 36(4), 365-375
- [3] Tesema, T. B., Abraham, A., & Grosan, C. (2005). "Rule mining and classification of road traffic accidents using adaptive regression trees". International Journal of Simulation, 6(10), 80-94.
- [4] Maze, T. H., Agarwai, M., & Burchett, G. (2006). "Whether weather matters to traffic demand, traffic safety, and traffic operations and flow".

