ISSN: 2321-2152

IJMECE

International Journal of modern electronics and communication engineering

E-Mail
editor.ijmece@gmail.com
editor@ijmece.com

www.ijmece.com

ISSN2321-2152

www.ijmece .com

Vol 12, Issue.2 April 2024

Improving Road Safety Using Machine Learning

Mrs.M.MamathaDevi [1]Jala Srinidhi[2]Y.Anurag Nandhan[3]Sai Kumar Ganta[4] Narlawar Srujan[5]

Assistant Professor/CSE(DS)TKR College of Engineering and TechnologyTelangana, India mmamathadevi@tkrcet.com

IV Final Year of CSE (DS)TKR College of Engineering and Technology Telangana, India Srinidhijala625@gmail.com

IV Final Year of CSE (DS) TKR College of Engineering and Technology Telangana, India Anuragnandan999@gmail.com

IV Final Year of CSE (DS) TKR College of Engineering and Technology Telangana, Indiasaikumarganta139@gmail.com

IV Final Year of CSE (DS) TKR College of Engineering and Technology Telangana, India narlawarsrujan25@gmail.com

---

ABSTRACT:

Road safety is one of the important actions which needs to be focused in the present world to avoid the damages caused. Accidents are most seen on the roads and many people are losing their lives because of these accidents and few of them were severely injured because of this accident. To reduce the number of accidents and to save the lives of people I am going to use machine learning techniques and big data concepts to improve the road safety. Machine learning is one of the commonly used methodologies for predicting the results with the data available and it is also used for visualizing the data properly and helps us in understanding the structured data or unstructured data easily.

Kew Words – Machine Learning, Big Data, Artificial Neural Network

1.INTRODUCTION

In the current world, Population has increased a lot and the needs of the people has also grown high when compared to olden days. Previously we used to travel more on the public transport rather than own vehicles but in the present situation the ideology of people has completely changed. As per the recent reports 77% of the households in UK has their own vehicles to travel and especially in England only 24% population doesn't have their vehicle and they are depending on the public transports like buses and tubes. Due to the high increase in the production of vehicles and usage of vehicles there are many chances of causing accidents which may lead to the death of the people or severely injured. As per the recent records of UK there were 24,470 people lost their lives or severely admitted in the hospitals because of severe accidents. Due to the covid pandemic situation and the lockdown condition the usage of vehicles has reduced in last year and because of which we can see a decrease in the accident rate and the death rate by 11% when compared to previous year. It clearly indicates that if we reduce the usage of own vehicles and utilize maximum percentage of the public transport then we can gradually decrease the death rate and the accident rate of the country. A country with less accident rate canbe considered as the safe.

2.PROBLEM STATEMENT

I have done some research on the reasons for causing accidents globally and across UK. I have decided to work on improving the road safety so that we can reduce the accident rate. To understand and analyze the accidents data I have chosen the UK accidents data available in the gov.uk website. In this data we don't have any personal information of the people. We only have the required data for analyzing the number of accidents caused and the factors

ISSN2321-2152

www.ijmece .com

Vol 12, Issue.2 April 2024

responsible for the accident caused. To accumulate the huge data, I have used big data concepts, for analyzing and predict the occurrence of accidents I have used Machine learning techniques like Xg boost and other methods. To achieve accurate results, I will be using the main concepts of big data in reducing the huge data to the required data so that we can train the model properly and the get the accurate results as we are removing all the noisy data from the data set.

3.PROPOSED APPROACH

After reviewing most of the research papers I have decided to use the most used algorithm which are suitable for the data set. To start the research, we need to decide a coding language a platform to execute the code written. I have chosen python as my base coding language. I have decided to use the UK data set released by the gov.uk. This data set is available in Kaggle, so I have extracted this data using Kaggle. For extracting the data set from Kaggle I have used google colab as my base platform for scripting language and we need to mount our google drive so that the extracted files will be saved in our google drive. As I am using big data concepts, I need huge amount of data, so I have gathered the maximum number of years data available from 2005 to 2014 which is 9 years accidents data. All these 9 years data are available in multiple files so I need to merge each individual file into one single file which can be used for the analysis part later. Each single file contains 33 features and the features or attributes for all the individual files are same. Using one primary key (unique feature) I have merged all the individual files to one file using pandas.

4.Testing Results

4.1 Data Analysis

From the below table we were able to get the count of number of causalities year wise and I have represented the same data using graphical representation. We can see that the number of causalities for the year 2012 has increased so I have done few analyses and identified that there are many duplicates available so tried to remove the duplicates. The count of the number rows in the year 2012 are 179715 with 34 columns in each row. After removing the duplicates and unnecessary data the count has reduced from 179715 to 90139.

From the below image we can observe that most of the accidents are happening in between 3:00 pm and 6:00 pm and least number of accidents happening in the morning time. With the help of weekly trend, we were able to observe that most of the accidents occurs on Friday and the highest number of accidents happened in the month of November and the least number of accidents happened in the month of February. With this analysis we were able to identify the features responsible for occurrence of accidents and the time, day and month were also identified. Now let's do some analysis on the area wise so that we can identify the highest accident-causing area and also, we can identify the same features for that particular area. In order to identify the area wise data, I have used the district code and the number of count of accidents happened in that area code or the district and arranged it in descending order so that we can get the highest count on the top.

	Year	Number_of_Casualties
0	2005	271017
1	2006	258404
2	2007	247780
3	2009	222146
4	2010	208648
5	2011	203950
6	2012	241954
7	2013	183670

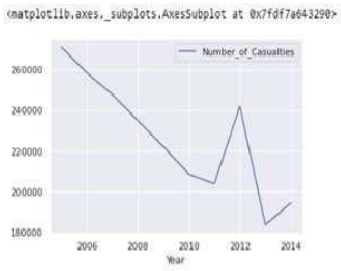
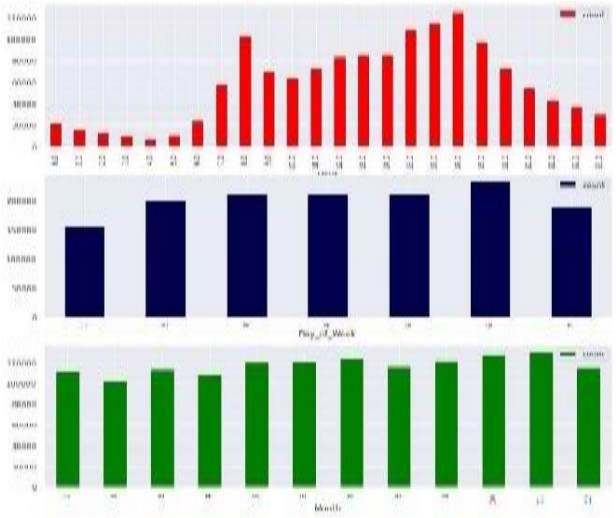


Figure 1: Causalities Trend



ISSN2321-2152

www.ijmece .com

Vol 12, Issue.2 April 2024

4.2 Models

4.2.1 Time Series Model:

In the time series from 2009 to 2014 we were able to get the rolling mean graph and also the de trending series graph.With the above graph it looks like the series are stationary but in order to check If the series or stationary or not we use ADF test which uses unit root test. For the time series consider the ADF test results are as follows :ADF Statistic: -5.793784 , p-value: 0.000000, Critical Values: 1%: -3.433 , 5%: -2.863,10%: -2.567 Here in this case the p value is greater than the 5% critical value, so the unit root test exits. If unit root test exits, then we consider this time series as not stationary and we fail to reject the null hypothesis. In order to get p value and q value we use ACF and PCF plots.

4.2.2Xg Boost Clasifier Model:

In this model I have trained the model initially and the actual fit vs the predicted fit was good, but I tried to tune the model for achieving the best results and the outcome of the model which was tuned has shown good results when compared to the initial model. The actual fit and the predicted fir for the final model was more accurate than the initial model. Initial model Scores before tuning:

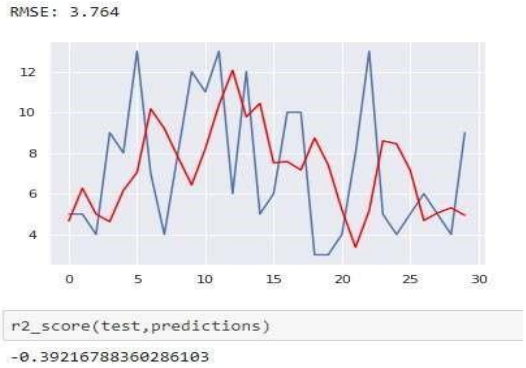
max_depth = 1100

MEAN RMSE: 2.138

MEAN R2: 0.519

Final model scores after tuning:

RMSE: 1.897 R2 score: 0.640



From the above figure 27 the blue color trend indicates the trend of the existing data, and the red color trend indicates that the prediction data. Whenever these two matches closely then we can consider that model as the best model.

5.DISCUSSION

From the data analysis part, we were able to see that the highest test number of accidents are occurring on Friday when compared to other days in a week. The severity of the accidents is also high on this Fridays. The maximum number of accidents are occurring in between 3:00 pm and 6:00 pm and in the early morning there are very less number of accidents happening. In the monthly trend the highest number of accidents occurred in the month of November and the least number of accidents occurred are in the month of February. When compared to all the districts the highest number of accidents occurred in the Birmingham city.

For Xg Boost Regressor in the initial prediction the results are as follows:

MEAN RMSE: 2.138

MEAN R2: 0.519

For the same Xg boost regressor model I have done tuning of the model and the results predicted after tuning are as follows:

RMSE: 1.897 R2 score: 0.640

Overall Result:

Model	RMSE value
Xg Boost Regressor after tuning	1.87
ARIMA model	2.10
Xg Boost Regressor before tuning	2.13
Random Forest	2.33
Fb Prophet Time Series	3.19

From the above table it clearly indicates that the Xg boost regressor model after tuning has shown the highest accuracy when compared to the other models used.

6.CONCLUSION

From this analysis I conclude that the Xg boost regressor model with fine tuning has predicted the proper results and the actual data trend vs the predicted trend is more accurate when compared to the other models. To achieve more accurate results for the other models we need to tune them properly so that there are many chances of getting more accurate values than the current values. With this I conclude that we need to focus mainly on Fridays as we have highest number of accidents occurring on this day when compared to other days. The second observation is that the time frame for a greater number of accidents are occurred in between 3:00 pm and 6:00 pm in the evening. At this point of time, we need to focus mainly as there is more flow of people happening on the roads as schools leave the students and most of the offices closes in this time interval only. The third observation for this research is that we need to focus mainly on Birmingham city as a greater number of accidents happened in this city only when compared to the other cities in United Kingdom.

As a future scope for this project, we can train the models with Artificial neural networks and we can get more accurate results. We can create a user interface which indicates the information which we have predicted in the maps using google maps API's so that the drivers can clearly see a pop up or can receive an alert stating that there are more chances of accidents occurring at this time interval and make the driver cautious so that we can reduce the number of accidents occurring in that particular interval of time. The alert system can be added to the police forces also so that they can also focus on the areas where we have highest number of accidents happening and also it will be easy for them to identify the features causing the accidents.

REFERENCE

[1] Silva, P.B., Andrade, M. and Ferreira, S., 2020. Machine learning applied to road safety modeling: a systematic literature review. Journal of traffic and transportation engineering (English edition).

[2] Tyagi, A., Kumar, A., Gandhi, A. and Mueller, K., 2019. Road accidents in the uk (analysis and visualization). arXiv preprint arXiv:1908.02122.

[3] Ljubič, P., Todorovski, L., Lavrač, N. and Bullas, J.C., 2002. Time-series analysis of uk traffic accident data. In Proceedings of the Fifth International Multi-conference Information Society (pp. 131-134).

[4] Yannis, G. and Karlaftis, M.G., 2010, January. Weather effects on daily traffic accidents and fatalities: a time series count data approach. In Proceedings of the 89th Annual Meeting of the Transportation Research Board (Vol. 10, p. 14).

[5] Lokala, U., Nowduri, S. and Sharma, P.K., 2017. Road Accidents Bigdata Mining and Visualization Using Support Vector Machines. World Academy of Science, Engineering and TechnologyInternational Journal of Computer and Systems Engineering, 10(8).

[6] Fiorentini, N. and Losa, M., 2020. Handling imbalanced data in road crash severity prediction by machine learning algorithms. Infrastructures, 5(7), p.61.