Analysis Of Right Turn Accidents in California

Akshay Anil Pagar
Nikita Bairagi
Sneha Thomas
Vandana Kudigrama Shenoy

akshayanil.pagar@sjsu.edu
sneha.thomas@sjsu.edu
nikita.bairagi@sjsu.edu
vandanakudigrama.shenoy@sjsu.edu

Advisor: Prof. Rakesh Ranjan

Department of Computer Engineering, San Jose State University, California

---

Abstract Right turn Accident is frequent problem for big vehicles. Accidents can occur for many reasons, such as when other vehicles squeeze into the are along the right side of your vehicle, failure to signal turn, motorists not recognizing turn signals, driver was distracted etc. There is a need to analyze what are the reasons for these accidents. Are they frequent in particular right turns and what causes them. In this project we will use the dataset for right turn accidents. Using data mining we will locate the accident prone spots and factors causing right turn accidents. By taking corrective measures on these spots based on analysis, number of accidents can be reduced.

Keywords ARIMA (Autoregressive Integrated Moving Average), SWIRTS (The Statewide Integrated Traffic Records System) , TIMS (Transportation injury mapping system)

---

1. INTRODUCTION

According to the Insurance Institute of Highway Safety (IIHS), the deaths occurring from motor accidents in USA is on the rise [1]. California is one of the state that tops the chart and accounts for a large share of these accident deaths. The population in California is increasing every year with the booming economy and attractive climate being some of the reasons. The increase in population results in more number of vehicles on the road and more accidents which is the reason why the rate of accidents in California is almost double the national rate. The Statewide Integrated Traffic Records System (SWIRTS) is a database managed by the California Highway Patrol where they record details about a collision. This open

data set can be used to analyze the patterns of right turn accidents in the state. According to the data, the deadliest counties are Los Angeles, San Bernardino, Riverside, San Diego, and Orange.

California is one among the many states in US where right turn at a red light is allowed after stopping unless they explicitly say so with a red

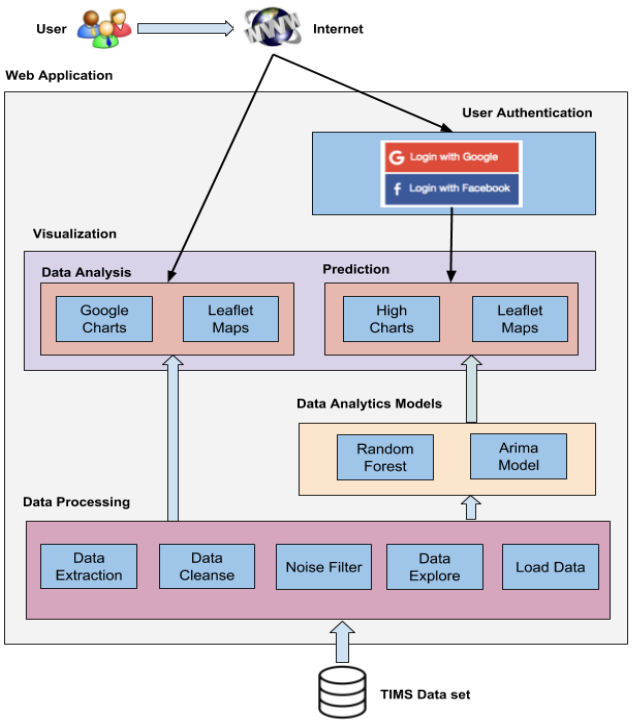
turn arrow or a sign board. It is the responsibility of the driver to make sure that no pedestrians are

crossing, no bicycles are there in the roadway and no vehicles are in the oncoming traffic. Besides the general causes for accidents, there are some common reasons that are seen more often in case of right turn accidents such as not stopping completely and the intersection, not checking for bicycles or motorcycles in the rearview mirror, not using blinkers while taking the turn, not checking the oncoming traffic properly. Apart from these reasons there are some other factors such as the weather conditions, condition of the road, and time of the day when accident occurred. If the weather conditions are bad such as foggy or rainy, the visibility of oncoming traffic, pedestrians, or bicycles at certain intersections can become difficult. If the road condition is bad or some construction is going on, and if the driver is unaware of this situation, it can cause difficulties in driving. If there are holes in the road it can cause the driver to lose control or if the road is flooded, it can make the road slippery and lose traction. The time of the day is another factor – the traffic is more usually from Morning 7.00AM-9:00 AM and in the Evening 03:00PM-7:00PM which can make the driving difficult. Another factor is the lighting at a time of the day – during the day with more sunlight, it increases the visibility of the road and driver can keep an eye on the bicycles, pedestrians or motorcycles around. This can get hard in the dark with less visibility.

The sad part of these accidents is that many of them could be avoided if there is a system to provide users with some information on the precautions that needs to be taken while travelling from location A to B, like to watch out for bicycles. Our web application "Right Turn Accident Analysis in California" has been implemented with an intention to provide its users with heads up of factors that can help them avoid right turn accidents based on various factors. When a user wishes to travel from one place to another, they can access our web application to identify various factors that can lead to a right turn accident along their route.

2. ARCHITECTURE

This web application is designed to provide users with analysis and prediction of accident probabilities based on California right turn accident data set. At first, the raw data is cleansed to obtain meaningful information out of it. This processed data is then used to provide data analysis and prediction. Machine learning prediction models which suits the problem type and data set are identified and implemented.



The Right Turn Accident Analysis in California web application consists of the following layers:

2.1 DATA SOURCE

For analysis and prediction, the data from TIMS – Transportation injury mapping system has been used which is based on SWIRTS. It contains extensive details related to all the collisions that

has happened in various counties of the state of California.

2.2 DATA PROCESSING LAYER

The raw data from the data source is processed in the back-end layer using Python to filter out noise and unimportant data. This data is then used for the data analysis and prediction. The various steps involved in the data processing layer includes the following:

a. *Data Extraction*: In this step, the data is extracted into the application from the data source
b. *Data Cleanse:* The raw data is then cleansed to remove unnecessary fields, add correct data types to the fields, remove redundant field etc. and only the relevant data is retained.
c. *Noise filter:* There can be some noise data with unsupported characters and spaces which needs to be removed.
d. *Data explore:* In this step, new fields are added, and data is segregated from various sources and provided meaning.
e. *Load data:* The processed data is then loaded in this step for the next layer to process.

2.3 DATA ANALYTICS MODEL LAYER

a. *Random Forest Model:* For the prediction of accident possibility at an area, classification model called Random Forest was used.
b. *ARIMA model:* The time series prediction of accident count was done using ARIMA model.

2.4 VISUALIZATION LAYER

a. *Data analysis:* The visualization of the data analysis is done using google charts and leaflet maps on a responsive web UI. Pie charts and heatmaps are added to help the user to deduct meaningful information swiftly. We often use maps for navigating to a location or look at a route to plan a travel. Getting an accident analysis data on a map helps making these plans safer and easier.
b. *Prediction:* For visualization of the prediction results, highcharts and leaflet maps has been used on a responsive web UI. After entering some factors such as weather condition, time of day, and road condition, users can get a prediction result of the precaution that needs to be taken while driving through a particular location. The prediction of time series provides a forecast of how the accident figures are going to be in the upcoming years for the authorities to take a note of.

2.5 USER AUTHENTICATION LAYER

If a user wishes to access the prediction screens, they need to be registered with the application. Google and facebook authentication has been provided for user to perform the login without much hassle. There is another option too to enter user details manually to get registered with the application.
For accessing the data analysis charts and maps, there is no need for the user to register with the application and the login is required only for accessing the prediction features.

3. DATA

The data used for this project was extracted from Transportation Injury Mapping System database from year 2006 to 2017. The accidents database includes different sets of data such as collision factors, party factors and victim factors. Table 1 summarizes the collected data.

Table 1: Data reported by Transportation Injury Mapping System

Data set	Collected Information
Collision Factors	County, city, collision severity, number killed, number injured, PCF violation category (Pedestrian right of way, Pedestrian violation etc), population, weather, lighting, road surface, road condition, pedestrian accident, bicycle accident, motorcycle accident, truck accident, latitude, longitude
Party Factors	Movement preceding collision - Right Turn

4. DATA ANALYSIS

In the application, we are providing 3 screens for data analysis:

4.1 HEAT MAP

A heat map is provided with the map of California to show the number of accidents. It gives a quick overview of the number of accidents that has happened in various counties with the shades of red and green. Light green shows less number of accidents and increases as the shade becomes more red. On hovering on top of the counties, it gives the number of accidents that has occurred in that county. To view details of accidents that has occurred in a particular county, user can click on it on the map which takes them to the second screen of data analysis with charts. The heat map has been implemented using leaflet maps.

4.2 CHARTS

An option is provided in the charts screen to select the county. A user can directly come to this view or can come through clicking on a county in the heat map. A summary of accidents about the county is provided in the top which categorizes the accident that has happened in that county. 2 pie charts have been provided in the charts screen to provide user with an analysis of the primary collision factor and the severity. These charts has been developed using google charts.

Collision percentage by primary collision factor: It shows the primary collision factors of accidents that has happened in a county such as improper turning, wrong side of road etc. It gives a clear picture to the user of what precaution needs to be taken while driving through a particular county

Collision percentage by collision severity: This pie chart helps the user to analyze the severity of accidents that has happened in a county. It categorizes the accidents like severe, complaint of pain etc.

4.3 COLLISION MAP

When a user plans a travel, it will be very helpful if we can provide them with the locations that are more accident prone. This map marks all the locations where accidents has occurred in a county. It helps user to identify areas where more number of accidents occur and take extra care for a safe travel. On changing the county, it clears and populates next set of data corresponding to the new county selected. This map has been implemented using leaflet maps

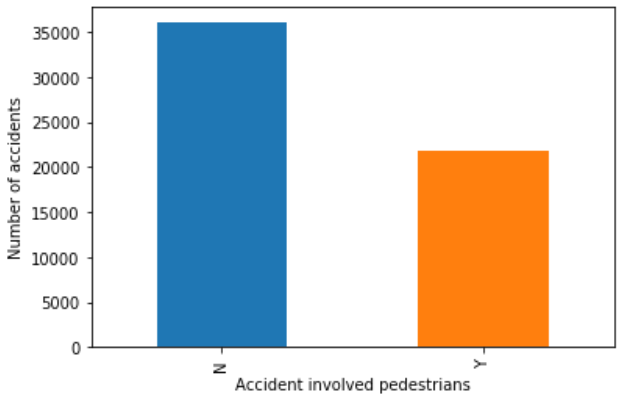
5. CLASSIFICATION MODEL

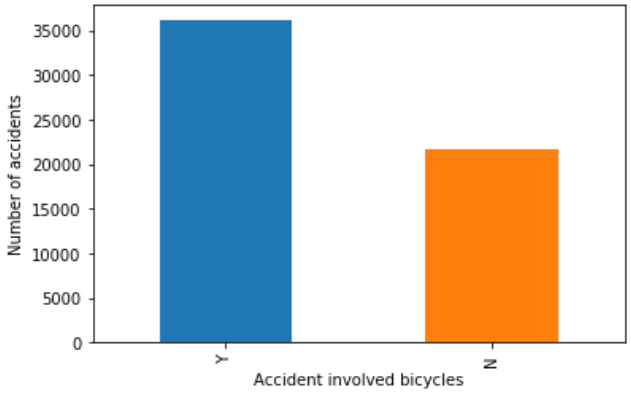
In this project we are trying to predict if there will be an accident involving pedestrians, trucks, bicycles or motor cycles at a given location. This will be based on factors such as weather condition, road condition, lighting at that location, time of the day, day of the week etc. We also evaluate the surrounding of the selected location like number of restaurants, schools, apartments etc. present. This will help us differentiate between different types of location such as

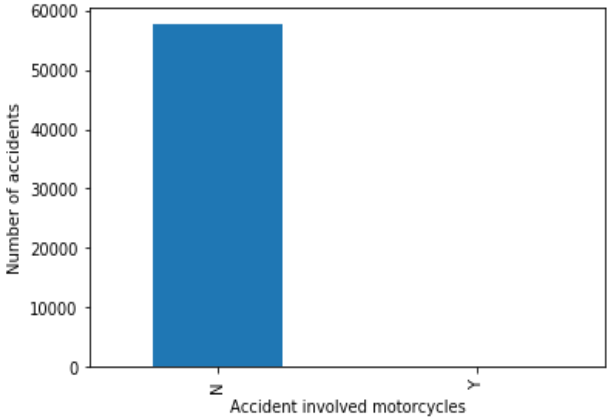
commercial or residential. By this, we will be able to decide which location will have more traffic.

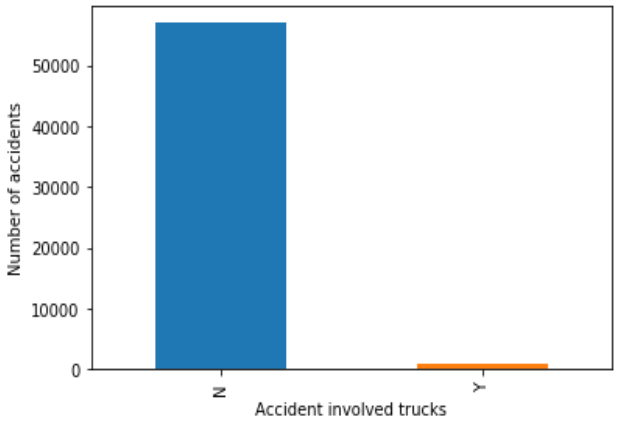
5.1 DATA DISTRIBUTION

Below are the graphs showing the distribution of accidents involving pedestrian, bicycles, motorcycles and trucks.









As we can see, the data is unbalanced for each of the classes. So, we are using F1-score to validate our model.

5.2 FEATURE ENGINEERING

5.2.1 HANDLING CATEGORICAL ATTRIBUTES

Categorical attributes such as weather condition, road condition, day of week etc was handled by creating one column for each of its value.

For example, for day of week we created 6 columns as mon, tue, wed, thur, fri and sat ('sun' would be a dependent attribute on these 6 attributes) , if we select a date such that the day of week is Monday the attribute "mon" will have value 1 and rest of the attributes will have value 0.

5.2.2 LOCATION INFORMATION

We got the latitude and longitude for a selected location and used 'HERE LOCATION SERVICES' [2] to get information of the surroundings like number of restaurants, schools, apartments, parking facilities, hospitals etc. present within 50 mts of the given location. We then create feature for each of this with values being their count. We create almost 80 such features based on surroundings.

For example, if there were 2 restaurants and an airport, the values for restaurants would be 2 and for airport would be 1.

5.3 METHODOLOGY

Random forests [3] or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. We have used Random Forest Classifier from scikit learn to make the predictions as the relationship between the attributes and labels is non-linear.

6. TIME SERIES FORECASTING

To further understand the trend of road accidents and road fatalities, we have used time series forecasting for predicting the number of accidents, number of injured and number of pedestrian injured. For prediction of number of total accidents, the granularity of the data was changed from per accident to total accidents each month using aggregation of data points. We were able to run prediction on this aggregated data filtered by location and forecasting is done for next one year to next five years. The data was fitted into an ARIMA (Autoregressive integrated moving average) model. For validating the algorithm we initially trained the model, by dividing 10 years of data in two parts, 7 years data for training and 3 years of data for testing accuracy. The root mean square error was found to be 46.34.
Following is the graph which represents prediction v/s the actual.

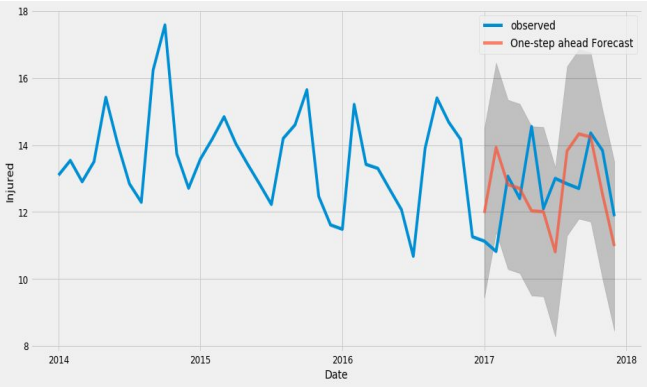


Fig: Actual Vs Prediction For next 1 year

Following is the next 3 years prediction graph for people injured in road accident based on last 10 years of data.

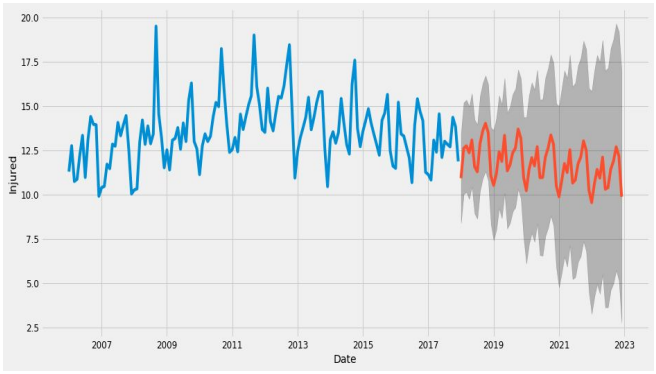


Fig: Prediction of next 3 years

The above predictions runs for total accidents per year by aggregating data for each month. This can also be extended to predict the accidents filtered on cause, and other classification columns.

7. CONCLUSION

This project focused on the severity of injuries and fatalities of pedestrians involved in the road collisions, especially when they were trying to cross roads. The impact of some external factors regarding road and site characteristics on the severity levels was investigated. The road factors analyzed included the road width and road type. The site factors included crash location and presence of pedestrian crossing. The employed data was extracted from TIMS database for the past ten years (from year 2007 to 2017). The analysis showed that the highest dangerous situation of pedestrian crossing takes place in case of the absence of crosswalk facility and drivers not giving pedestrian right of way. It is expected that the findings of this study will help traffic engineers and urban planners to implement more effective countermeasures for pedestrian crashes as well as for developing a criteria for selecting the appropriate crosswalk facilities.

9. REFERENCES

[1]https://www.iihs.org/iihs/topics/t/general-statistics/fatalityfacts/overview-of-fatality-facts
[2]https://www.here.com/en/products-services/here-location-suite/here-location-services/overview
[3]https://en.wikipedia.org/wiki/Random_forest