**US-Accidents: Predicting and analyzing countrywide road accidents**

Team Members: Abhilash Sridhar, Harshit Sanwal, Prateek Gulati, Shashank Shekhar, Zalon Wong

## Introduction

The United States is one of the busiest countries in terms of **road traffic** with nearly **280 million vehicles** in operation and more than **227.5 million** drivers holding a valid **driving license in 2020**. Some **6.7 million** passenger cars were involved in U.S. traffic crashes.

Every year **1.25 million people die** due to road accidents. The number of people getting seriously injured or disabled is 20–50 million annually. Among causes of death, traffic accidents rank 9th and cost US $231 billion every year.

The traffic-related **fatalities per one million** population for the United States stands at **1.13** which is among the highest in the world. There is a huge cost associated with the high fatality rate as evident from the liability due to auto insurance losses in the U.S, amounting to approximately **$96.2 billion**.

The average auto **insurance expenditure** in the United States for 25-year-old driver is approximately **$3,207 per year** for car insurance which is one of the highest in the world.

Increasing traffic every quarter, lax driving tests and lack of training are sometimes sighted as reasons for the high level of road accidents and fatalities in the United States, compared to other countries. Statistics say that male drivers are behind the wheel in majority of fatal crashes. Speeding and driving under the influence of alcohol have often been to blame but neither have led to a rise in traffic fatalities in the U.S. Instead, cellphones are likely the primary cause as drivers are distracted by texts and social media. Young drivers are the ones accounting for the highest share of cellphone use fatalities.

In this project, we try to generate actionable insights to reduce the number of accidents by identifying various patterns and key factors contributing to road accidents in the United States.

Dataset Description

Dataset is published on Kaggle.com. The data was collected from 3 sources, namely MapQuest and MapQuest-Bing, and Bing. The dataset contains more than **4 million records** with **49 variables** with as latest as accident reported on **Dec,31,2020**. Below is a snapshot of Data exploration. The variables have been categorized as nominal, ordinal or numerical variables. Variables have also been determined as redundant or not.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Groupings** | **#** | **Column Name** | **Type of Variable** | **Redundant Variable** | **Missing Values** | **Data exploration remarks** |
| Accident | 1 | ID | Nominal | Yes | No | Unique values and the total values are same, suggesting **no duplicate records**.  Should be removed.  Does not provide any insights w.r.t severity of accidents. |
| 2 | Source | Nominal | No | No | Definitions of Severity of accident are different for different Sources. MapQuest: 63% Bing: 36% Other: 2% |
| 3 | TMC | Nominal | No | Yes | 36% of the values are missing. |
| 4 | Severity | Ordinal | No | No | Severity-based analysis per 'Source'. As different sources have varying criteria for severity. (Assumption: Severity from different sources is classified on similar parameters) |
| 5 | Start\_Time | Numerical | No | No | Format:  “yyyy-mm-dd hh:mm:ss”  Example:  "2020-12-31 23:28:56". Should be used to calculate **derived variable “Duration”.** |
| 6 | End\_Time | Numerical | No | No | Should be used to calculate **derived variable “Duration”.** |
| Geographical Factor | 7 | Start\_Lat | Numerical | No | No | Useful for descriptive analytics. |
| 8 | Start\_Lng | Numerical | No | No | Useful for descriptive analytics. |
| 9 | End\_Lat | Numerical | Yes | Yes | Approx. 64% of observations have null values.  Should be removed.  Does not provide any insights w.r.t severity of accidents.  Even if we impute these variables with Start\_Lat for those with zero distance, still most of the observations will be same to Start\_Lat (As majority of the distance involved in the accidents is 0 miles).  Making them highly correlated. |
| 10 | End\_Lng | Numerical | Yes | Yes | Same as End\_Lat (#9) |
| 11 | Distance mi | Numerical | No | No |  |
| 12 | Description | Nominal | No | Yes | Not directly related in predicting the severity of accidents.  Word Cloud generated out of this data for meaningful insights. |
| 13 | Number | Numerical | Yes | Yes | Does not provide any insights w.r.t severity of accidents. |
| 14 | Street | Nominal | No | No | Does not provide any insights w.r.t severity of accidents.  Won't be used in regression analysis, however, useful in descriptive analytics. |
| 15 | Side | Nominal | No | No | (Not Sure) |
| 16 | City | Nominal | Yes | Yes | Does not provide any insights w.r.t severity of accidents.  Won't be used in regression analysis, however, useful in descriptive analytics. |
| 17 | County | Nominal | Yes | No | Does not provide any insights w.r.t severity of accidents.  Won't be used in regression analysis, however, useful in descriptive analytics. |
| 18 | State | Nominal | Yes | No | Does not provide any insights w.r.t severity of accidents.  Won't be used in regression analysis, however, useful in descriptive analytics. |
| 19 | Zipcode | Numerical | Yes | Yes | Does not provide any insights w.r.t severity of accidents.  Won't be used in regression analysis, however, useful in descriptive analytics. |
| 20 | Country | Nominal | Yes | No | Does not provide any insights w.r.t severity of accidents. |
| Time & Weather | 21 | Timzone | Nominal | Yes | Yes | Does not provide any insights w.r.t severity of accidents. |
| 22 | Airport\_Code | Nominal | Yes | Yes | Does not provide any insights w.r.t severity of accidents. |
| 23 | Weather Timestamp | Numerical | Yes | Yes | Does not provide any insights w.r.t severity of accidents. |
| 24 | Temperature F | Numerical | No | Yes |  |
| 25 | Wind\_Chill F | Numerical | Yes | Yes | Highly correlated with Temperature.  Approx. 46% missing values.  Should be removed. |
| 26 | Humidity | Numerical | No | Yes | Approx. 2% missing values. Either impute county wise average value or drop the observations. |
| 27 | Pressure | Numerical | No | Yes | Approx. 2% missing values. Either impute county wise average value or drop the observations. |
| 28 | Visibility | Numerical | No | Yes | Approx. 2% missing values. Either impute county wise average value or drop the observations. |
| 29 | Wind\_Direction | Nominal | No | Yes | Approx. 2% missing values. Either impute county wise average value or drop the observations. |
| 30 | Wind\_Speed | Numerical | Yes | Yes | Vehicle speed is required to get relative speed which would be a better predictor.  Range of values 0-18mph Mean of 8mph  The above wind speed levels suggest that this variable will not impact much. |
| 31 | Precipitation in | Numerical | No | Yes | Approx. 48% missing values. Impute county wise average value. |
| 32 | Weather\_Conditions | Nominal | No | Yes | Convert into a Numerical variable with no order. |
| Road Conditions | 33 | Amenity | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 34 | Bump | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 35 | Crossing | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 36 | Give\_Way | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 37 | Junction | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 38 | No\_Exit | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 39 | Railway | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 40 | Roundabout | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 41 | Station | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 42 | Stop | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 43 | Traffic\_Calming | Nominal | No | No | Contains True/False values.  Only one dummy binary variable to be created. |
| 44 | Traffic\_Signal | Nominal | No | No | Contains True/False values.    Only one dummy binary variable to be created. |
| 45 | Turning Loop | Nominal | Yes | No | Contains only False values. |
| Day Light | 46 | Sunrise\_Sunset | Nominal | No | Yes | Have the same info as Civil\_Twilight.  Should be removed. |
| 47 | Civil\_Twilight | Nominal | No | Yes | 0.003% missing values.  Majorly has two values “Day” and “Night”.  Need to impute with the value as per the Start\_Time.  However even if we impute with any of the value it will not impact much because of very low frequency. |
| 48 | Nautical\_Twilight | Nominal | No | Yes | Have the same info as Civil\_Twilight.  Should be removed. |
| 49 | Astronical\_Twilight | Nominal | No | Yes | Have the same info as Civil\_Twilight.  Should be removed. |

**Section 3: Descriptive Analytics**

Chart, bar chart

Description automatically generated

**Frequency of Accidents w.r.t Severity**

Majority of the accidents are not severe. Most of the accidents are of Severity level ‘2’.

Chart

Description automatically generated

**Accident Severity**

**City vs Highway**

Less severe accidents happen in the city area while more sever accidents take place on highways

Chart, bar chart

Description automatically generated

**Accident vs Month**

Accidents mainly happen in the month of Dec or the decond half of the year

Chart, bar chart

Description automatically generated

**Accident vs Day of the Week**

Most of the accidents are happening over Weekdays

Chart, histogram

Description automatically generated

**Accident vs Hour of Day**

Majority of the accidents are happening in the morning time between 7am to 8 am followed by in the evening between 4pm and 6 pm, clearly indicating office hours

Chart, histogram

Description automatically generated

**Accident vs Temperature**

Most accidents are happening when under ambient temperature, ranges from 60-80 degrees Fahrenheit

Map

Description automatically generated

Most of the accidents happening on east coast and west coast

Chart

Description automatically generated

**Top 10 states w.r.t Accidents**

California contributes majorly to accidents followed by Texas and Florida.

Chart, bar chart

Description automatically generated

**Top 10 Cities w.r.t Accidents**

Houston contributes majorly to accidents followed by Charlotte and Los Angeles.

![A picture containing logo

Description automatically generated]()

**Word Cloud**

Text mining of the description of accidents suggest that the accidents happen primarily because of lane change and exit

## Analysis and Way Forward

Goals of Analysis

Based on our exploratory data analysis, we identified two variables of interest, namely ‘Severity’ and ‘Duration of Impact’ (reflected by the difference between ‘Start Time’ of the accident and ‘End Time’ of the impact from the original dataset).

The goal of this project is to identify factors pertaining to geographical locations, time and weather, and road conditions that are associated with accidents in the United States. We will look into the three categories of factors and figuring out the ones with highest correlation with accidents. Apart from a national investigation, we will also zoom into states and/ or cities to study the patterns of accidents. For example, we will look at whether the time and weather; and road conditions factors have different relationships with accidents in states with highest number of accidents in the past four years.

Being able to predict the severity and impact of accidents; and understand the major drivers of accidents are not only beneficial for accident prevention but also mitigation of the damage caused by accidents. There are many use cases of the potential knowledge generated from this project. For instance, from a city planning perspective, the conclusion of the project will shed light on safe road designs. The algorithm derived from this project could be useful to make accurate prediction of impact duration of accidents. If such information can be broadcasted through radio, GoogleMaps and other popular applications, this will save the general public from traffic jams.

Techniques and Methods

Given the variety of variables, we expect make a huge effort in feature selection. Before adopting any specific measurement tools, we assessed the feasibility of using the variables available as predictor variables based on our understanding of the dataset as well as the techniques that we will use. We determine that some variables can be excluded from the regression analysis given the information they carry and the complexity of the variable. For example, ZIP code will be eliminated as it will add tremendous complexity to the regression analysis and it has overlapping information with other geographical variables.

After making sensible judgement of the features, we will rely on various statistical measures such as correlation coefficients and Variance Inflation Factor (VIF) to evaluate multi-collinearity. Pearson correlation coefficients can be found among numeric variables. After generating a base regression model, we will examine VIF and Cook’s distance to rule highly correlated variables and spot outliers.

Model building and feature selection is a reiterative process, especially for a big dataset like ours. Residual analysis will be run to check whether the basic assumptions of regression are met and further identify outliers. Normality Probability Plot, Residual Against Fitted Values Plot, and PRESS Residuals will come in place of assessment at this stage. Cook’s distance can also be employed to identify influential data point (i.e., outliers). In case of violation of regression assumptions, we will further explore the need of variable transformations such as standardization and Box-Cox transformation. In the meantime, outliers will be removed.

A good regression model is the one that captures significant factors while minimizing the number of regressors involved. To come up with a good regression model, our decision-making will rely on several statistical measures including Adjusted Coefficient of Multiple Determination (Adjusted R2), Akaike’s Information Criterion (AIC), PRESS Criterion and Mallow’s Criterion. These measures will be tracked and compared as we modify the model through stepwise regression procedures.

Having selected a few candidate models, we will enter the model validation stage. With an abundance of observations (~4 millions), we will run a 10-fold cross validation to improve model accuracy. Given the large size of dataset, it is possible to re-run 10-fold cross validation more than once to improve our confidence of the models. Apart from the statistical measures mentioned above, we will also generate confusion matrices enhance our understanding of the performance by looking at accuracy, error rate, sensitivity and specificity. By comparing the performance measures of the training and testing datasets, we can also evaluate the overall goodness of fit of our models.

To conclude, the feature selection and model validation processes are highly intertwined, and we will be conducting both back and forth before a desirable model is concluded.

Comments from Prof.

Narrow down to city, philly.

Check the number of zip codes in philly

Add all the points from HW topics into the project.

Duration: