# Title : Findings from Analysis of Coursera Capstone Project California State Collision Data.

Prepared by : Periyapatna V Chandrashekara

Designation : Global Offering Manager @ IBM Corporation.

Published : 10 Sept 2020

# Table of Contents :

1. Introduction
2. Data Sources and Methodology of Data analysis
3. Results of Data Analysis
4. Discussion Forum
5. Conclusions
6. References
7. Acknowledgement

# Introduction

The work done in this project is analysis of the Accidents (Collision Data)  from state of California USA. This report contains  findings and recommendations that help  control and reduce the number of collision Incidents. And a guidance to creation of a system that provide proactive notification and warning to  public on  road safety and driving conditions, and precautionary  steps to observe and follow guidance by the  State Transport Authority.

**Problem Statement** : The data from the source  indicates very large number of collision incidents, resulting in injuries   higher in numbers. The number of accidents is consistently high with average of about 12000 incidents per year resulting  in road safety concerns and damage to personal health and property.

This report provide you details of  conditions and factors responsible for  collision incidents and  methodology used in this research  to  prepare  the findings and  recommendations.

**From the data,** the cause of collision incidents cover range of samples  like  pedestrians, animals crossing,parked vehicles, vehicles in motion being involved in accidents. The weather conditions like rain, snow , road conditions like icy, slushy, and lighting conditions like Dark (night) with or without street lights, foggy etc being recorded as they contribute significantly to the cause of collision incidents..

Also  driver behavior ( like speeding, under the influence ) are also cause of accidents,  again are augmented by the weather/ lighting conditions.  Most of them occurring  during night when street lights are on.

While the existing traffic Law have defined guidelines for drivers to observe/follow rules  to prevent collision incidents ( ex: speeding, under influence etc),  preventing incidents as a result of weather/road condition requires new specific measures and guidelines.

The data analysis will provide insights and remediation suggestions

Research  Methodology

The data source  is the data file provided by Coursera  learning team.  Data cleansing is done as data being inconsistently updated for each incident and only relevant variables influencing cause of the incident are considered.   Also data with unknown or “other” as categorical values are ignored as they do not contribute to learning knowledge.

With Target variable being the severity of incidents ,  data modeling is considered for predicting likeliness of collisions ending up with injuries due to influence of weather, Road an light conditions. SVM and Decision Tree techniques are applied to leverage the best model to gain high accuracy of prediction. Following steps taken to process the data, modeling and visualization.

1.   Statistical/descriptive analysis on the data set.

2.   Data cleansing and creation of pre-processed data set for modelling

3.   A correlation map of the variables resulting/causing collision

4.   Data modelling with Decision Tree  and SVM

5.   Prediction and model accuracy evaluation

6.   Visualization of data

# Discussion forum

The data contains information pertaining to each collision , however missing data is significantly high in many fields, are not considered (not being cause of collision).    For the critical variables, missing data rows are removed, as the data set is significantly large and number of missing data in the final set are minimal).   Speeding  and  UnderInf ( under influence) variables are not considered based in the ratio of data units (incidents).

From the statistical analysis we find, there are only 2 type of  result,  injury or property damage.  Hence this is a classic  case for categorical data analysis ( like the churn).

The data contains  location of accidents  in variables like  co-ordinates, Junction type( both as intersection  and  block), street /block address .  from the analysis we find there are few streets with a greater number of collisions. We will plot the data to provide insights. However these variables are not critical influencing factors for cause of collisions. And hence ignored.  

Data set also contains  conditions like road condition, weather and  lighting ( day or lights on) which play critical role in cause of collision.  **These are critical variables influencing cause and also augmenting driver’s condition ( Speeding, under the influence) to result in collisions.**

Data set includes other possible causes like  Speeding,  or under the influence .However the  sighting of such is  relatively small in number

* Speeding ~9759 of all incidents
* under influence  ~9200 of all incidents

Assuming these are  possibly well-known causes already considered under traffic rules and law, will be ignored from the data set.

Another critical  variable is the collision description, where in  accidents with parked vehicles are quite handful. There are about ~ 23% ( 44424 of all cases)  **involving  parked vehicles**.  And also  accidents with  **one car stopped,** and other hitting at rear  or side is quite considerable ( ~287001cases) 14% ( in same direction)

* This would call in for a specific recommendation to consider remediation to  avoid these cases. Like review Parking guidelines/ space  on the streets.
* Many of incidents of later type are due to speeding or influenced by weather conditions/ road conditions etc.

So Our data set will contain  limited fields/ columns **:   severity, road, weather, lighting conditions.** will work on the data model and code based on this descriptive analysis. Bar Graphs reflecting influence of each of the categorical values on collision incidents help compare the correlation and insights to make strategy/recommendation for control and prevention of injuries (from collisions) and collision incidents

# Results

The source data set provide a view of severity 2 “Injury collision “ which about 30%. The bar graphs reflect about 20% of all collision are influenced by weather (Rain) and road conditions being Wet.

Also poor light conditions, specially at Dark (nights), the number of collisions are 25%,

~70% of all collisions have occurred during daylight and cleat sky conditions, reflecting other factors like(possible) driver attention to road (texting /using mobi1e/speeding) etc

Also, about 23% of all collisions are with parked vehicle, indicating the obstruction to the moving vehicles on the road.

The trend of year-on-year count of incidents are declining, helps the Authorities to continue the containment laws and efforts in place

# Conclusion

Based on the insights and results the following set of recommendations for action are suggested.

1. Review the parking conditions and adopt alternate parking / bring in “no parking” rules to contain number of collision incidents.
2. Clear “marking” /” sign boards and improved lighting at Driveway intersections in the block, and at the intersections to be ensured
3. A system to notify vehicle drivers/ owners on the road about road conditions, Rain conditions to be alerted ( develop a system for notification on the car dash board system, as well apply “Speed breaking” controls in the car when weather conditions are not right for vehicular movement.
4. Number of incidents resulting (%age) in injuries would decline per the prediction and accuracy results from the data model techniques.

# References

1. Pandas documentation version 1.2
2. Reviews , Q&A from Stack Overflow community
3. Modules and videos from Coursera course materials for “IBM Data Science Professional” certification

# Acknowledgements:

Thanks to Suraj Shekar, my son who just completed his under Graduation in Engineering in IoT. And his tips in python and data analysis helped me during the project.

My peers on this learning course, who have provided reviews and feedback on multiple assignments to better understand the core concepts and hands on experience on Data Science.

Thank you “The Instructors” of all the modules of this high value certification learning course. Your method and coverage of topics was very helpful in learning the subject more easily.