Analysis and Modeling of Car Crash Data in Chicago

By: Daniel Wahome

INTRODUCTION

CHICAGO is the most populous city in the U.S. state of Illinois and in the Midwestern United States. With a population of 2,746,388, as of the 2020 census, it is the third-most populous city in the United States after New York City and Los Angeles. As the seat of Cook County, the second-most populous county in the U.S., Chicago is the center of the Chicago metropolitan area, often colloquially called "Chicagoland" and home to 9.6 million residents.

BUSSINESS UNDERSTANDING

The City of Chicago has implemented E-Crash a reporting tool for car accidents that happen within the City of Chicago, but with all this recorded data they have not been able to make strides in minimizing accidents on the streets of Chicago city.

With the data recorded through the E-Crash system and use of machine learning principles, I can find primary contributory causes of car accidents and recommend solutions based on interpretation of results.

This will in turn help keep the citizens of Chicago safe through preventative measures therefore reduced car crashes, help in implementation of city planning principles to reduces the crashes and limit the amount of resources the City of Chicago utilises to solve these cases

OBJECTIVE:

Build a classifier to predict the primary contributory cause of a car accident, given information about the car, the people in the car, the road conditions and other various factors included in the data set

- **DATA UNDERSTANDING**
- The data source is the Traffic Crashes - Crashes dataset form the Chicago Data Portal and this data is suited to this task as it collected by the Chicago Police Department officers who are skilled about law enforcement and also that the citizens can contest the content of the data recorded and amendments made upon confirmation
- The data collected contains multiple characteristics like: speed limit, weather conditions, lighting conditions, crash type among many other characterizes that help paint a picture of how the car crash occurred. this data is suitable for our project as it provides comprehensive and relevant information needed to predict the primary contributory cause of car accidents, which aligns with our business objective to improve traffic safety.

- **TARGET VARIABLE**
- MOST_SEVERE_INJURY: This seems to be the target variable.
- It likely describes the most severe injury sustained in the crash.

The data set has 48 columns namely: ['CRASH_DATE', 'POSTED_SPEED_LIMIT', 'TRAFFIC_CONTROL_DEVICE', 'DEVICE_CONDITION', 'WEATHER_CONDITION', 'LIGHTING_CONDITION', 'FIRST_CRASH_TYPE', 'TRAFFICWAY_TYPE', 'LANE_CNT', 'ALIGNMENT', 'ROADWAY_SURFACE_COND', 'ROAD_DEFECT', 'REPORT_TYPE', 'CRASH_TYPE','INTERSECTION_RELATED_I', 'NOT_RIGHT_OF_WAY_I', 'HIT_AND_RUN_I','DAMAGE', 'PRIM_CONTRIBUTORY_CAUSE', 'SEC_CONTRIBUTORY_CAUSE','STREET_NAME', 'NUM_UNITS', 'MOST_SEVERE_INJURY', 'INJURIES_TOTAL','INJURIES_FATAL', 'INJURIES_INCAPACITATING','INJURIES_NON_INCAPACITATING', 'INJURIES_REPORTED_NOT_EVIDENT','INJURIES_NO_INDICATION', 'INJURIES_UNKNOWN', 'CRASH_HOUR', 'CRASH_DAY_OF_WEEK', 'CRASH_MONTH'

The data set has 835407 entries

- DATA CLEANING
- Various methods were used to do data cleaning
  - Unnecessary columns were dropped
  - Change if data types to for proper analysis
  - Missing values and duplicates handled appropriately and data visualized

EDA

Through EDA this is what I was able to find

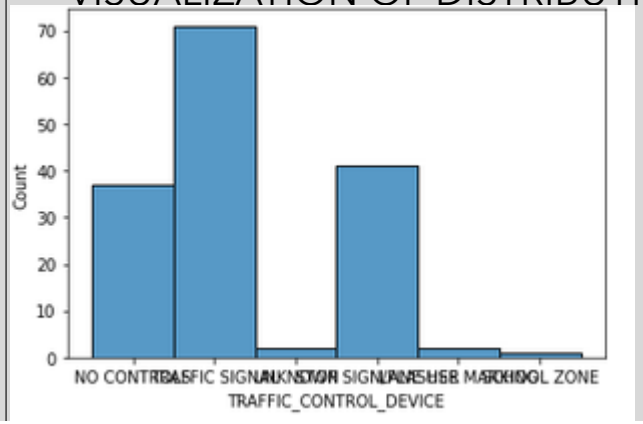
it seems that there might be a cause for most accidents to average around 11 pm on the 4th day of the week in July which also happens to be around The 4th of July celebrations,
we can also see that the most of the injuries are non-fatal

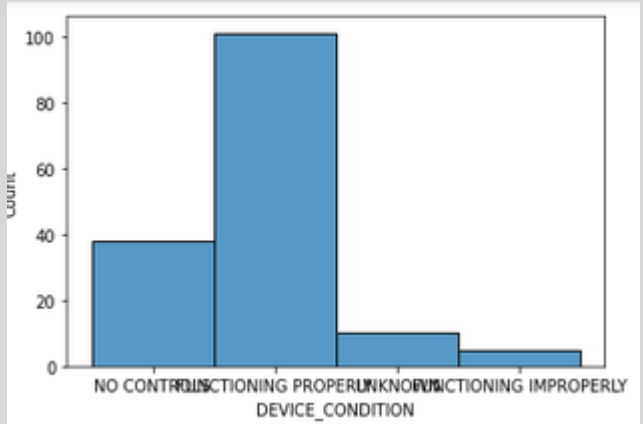
this shows that most accidents happen: during the day with clear weather, with traffic signals present, the roads are in good condition and dry accidents happen due to the driver hitting a fixed object and ending up with minimal injury but the car needing to be towed most damage ends up being at 1500 usd most accidents happen at cicero drive,
a mainstream, bustling and liberal part of chicago home to large latinx community

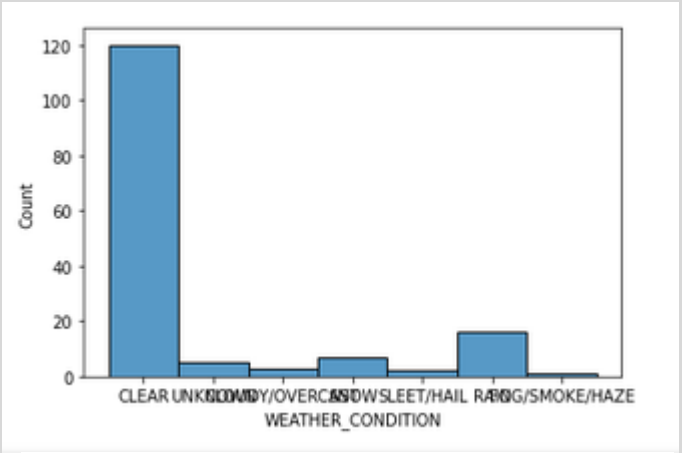
	count	unique	top	freq
TRAFFIC_CONTROL_DEVICE	154	6	TRAFFIC SIGNAL	71
DEVICE_CONDITION	154	4	FUNCTIONING PROPERLY	101
WEATHER_CONDITION	154	7	CLEAR	120
LIGHTING_CONDITION	154	6	DAYLIGHT	71
FIRST_CRASH_TYPE	154	12	FIXED OBJECT	44
TRAFFICWAY_TYPE	154	11	NOT DIVIDED	74
ALIGNMENT	154	3	STRAIGHT AND LEVEL	149
ROADWAY_SURFACE_COND	154	6	DRY	107
ROAD_DEFECT	154	7	NO DEFECTS	129
REPORT_TYPE	154	2	ON SCENE	110
CRASH_TYPE	154	2	INJURY AND / OR TOW DUE TO CRASH	97
INTERSECTION_RELATED_I	154	2	Y	136
NOT_RIGHT_OF_WAY_I	154	2	Y	100
HIT_AND_RUN_I	154	2	Y	130
DAMAGE	154	3	OVER $1,500	118
PRIM_CONTRIBUTORY_CAUSE	154	18	UNABLE TO DETERMINE	42
SEC_CONTRIBUTORY_CAUSE	154	21	UNABLE TO DETERMINE	47
STREET_NAME	154	99	CICERO AVE	6
MOST_SEVERE_INJURY	154	5	NO INDICATION OF INJURY	121

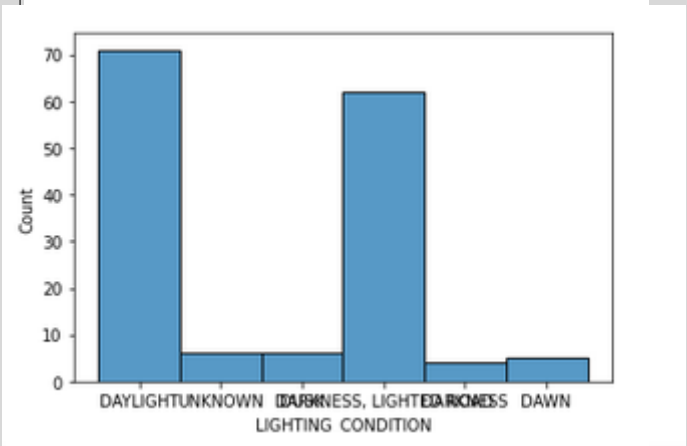
	count	mean	std	min	25%	50%	75%	max
POSTED_SPEED_LIMIT	154.0	29.415584	4.677440	0.0	30.0	30.0	30.0	45.0
LANE_CNT	154.0	2.558442	1.495033	0.0	2.0	2.0	4.0	7.0
NUM_UNITS	154.0	1.967532	0.931783	1.0	1.0	2.0	2.0	8.0
INJURIES_TOTAL	154.0	0.357143	0.954375	0.0	0.0	0.0	0.0	9.0
INJURIES_FATAL	154.0	0.012987	0.113588	0.0	0.0	0.0	0.0	1.0
INJURIES_INCAPACITATING	154.0	0.038961	0.252649	0.0	0.0	0.0	0.0	2.0
INJURIES_NON_INCAPACITATING	154.0	0.194805	0.616525	0.0	0.0	0.0	0.0	5.0
INJURIES_REPORTED_NOT_EVIDENT	154.0	0.110390	0.405222	0.0	0.0	0.0	0.0	3.0
INJURIES_NO_INDICATION	154.0	1.980519	1.411766	0.0	1.0	2.0	2.0	9.0
INJURIES_UNKNOWN	154.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
CRASH_HOUR	154.0	11.714286	6.971196	0.0	6.0	12.0	17.0	23.0
CRASH_DAY_OF_WEEK	154.0	4.064935	2.005467	1.0	2.0	4.0	6.0	7.0
CRASH_MONTH	154.0	6.448052	3.838304	1.0	3.0	7.0	10.0	12.0

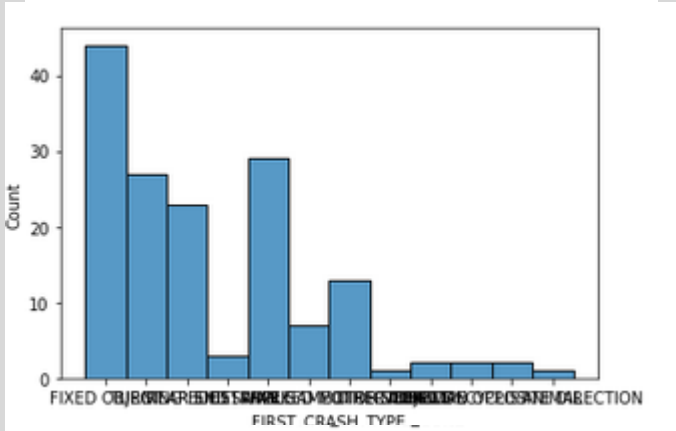
- VISUALIZATION OF DISTRIBUTIONS

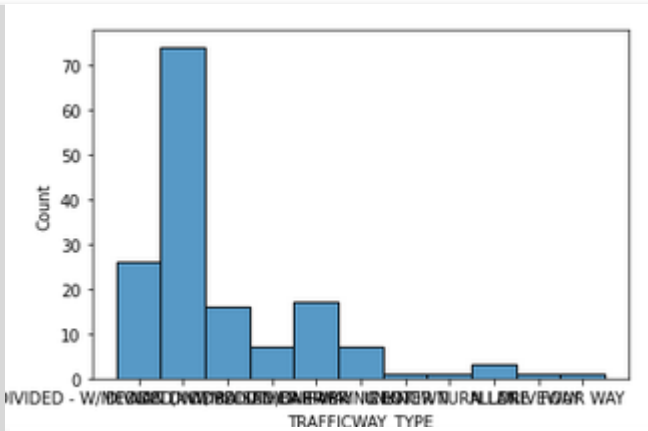


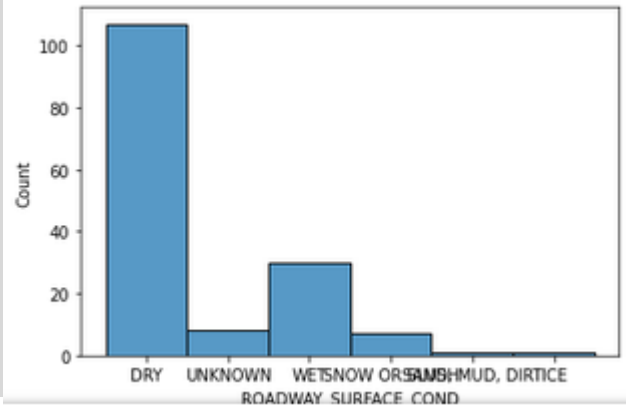


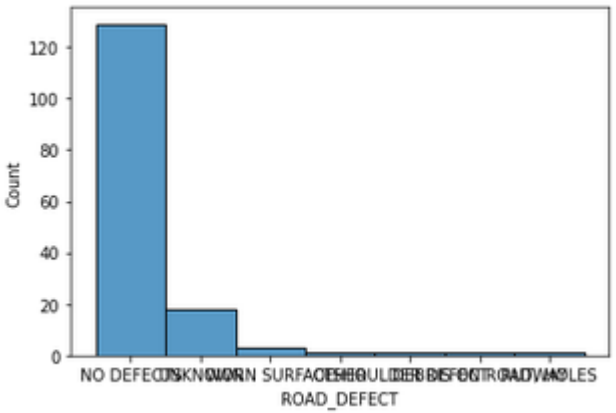


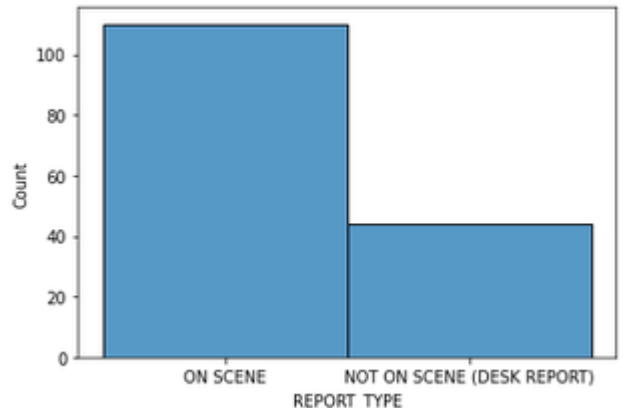












- univariate, bivariate and multivariate anaylsis was done, after eda was complete here are the summary of findings
- TIME Patterns: Most accidents occur during the day, especially between 6 AM and 6 PM. Additionally, crashes predominantly happen from Monday to Saturday, with March to October witnessing the highest frequency of accidents.
- Weather and Road Conditions: Clear weather conditions contribute to the majority of accidents, followed closely by rainy weather. The roads are typically in good condition and dry at the time of the crash.
- Location and Traffic Control: The majority of accidents occur in mainstream areas like Cicero Drive, where traffic signals are present. However, crashes also happen in locations with no traffic controls, indicating potential non-compliance with traffic regulations.
- Primary Contributory Causes: Primary contributory causes for accidents are often undetermined, followed by failure to yield the right of way and careless driving.
- Severity of Injuries: The most common type of injury reported is no indication of injury, indicating that many accidents result in minor damage. Fatal injuries are less common, with most crashes resulting in non-incapacitating injuries.
- Time of Day and Crash Types: Accidents occurring during the day are more likely to involve rear-end collisions and pedestrians, while nighttime accidents have a similar distribution of fatalities. Hit-and-run scenarios are more prevalent in fatal crashes.
- Seasonal Trends: There appears to be a spike in accidents around July, coinciding with the 4th of July celebrations. Despite the increase in accidents, most injuries during this time are non-fatal.
- Overall, the data suggests that most accidents in Chicago occur during daylight hours, under clear weather conditions, at locations with traffic signals. While many accidents result in minimal damage, there is a notable spike in crashes around July, potentially due to holiday celebrations. Understanding these patterns can aid in the development of targeted interventions to reduce the frequency and severity of accidents in the city.

- **MODELLING**

- PREPROCESS ANE NCODING WERE DONE HERE ARE THE RESULTS FROM MODELLING

THE FIRST REGRESSION MODEL OVERFITTED WHEN IT CAME TO ACCURACY WITH 100% ACCURACY AND HAD A VALIDATION SCORE OF 72%

THE SECOND CONFUSION MATRIX GAVE A BETTER PREDICTION WHEN IT COMES TO MORE NON FATAL CRASHES HAPENNING AT 85% BUT NOT ON HOW TO GREATLY REDUCE FATAL ONES AT 74% OVERALL

THE THIRD REGRESSION MODEL TARGETING THE RLSHP BETWEEN CRASH TYPE AND WEATHER CONDITIONS THE ACCURACY WAS REALLY LOW AT 29% THEREFOR NOT MUCH COULD BE PREDICTED FROM IT

THE FOURTH MODEL A DECION TREE TARGETING THE RLSHP BETWEEN CRASH TYPE AND WEATHER CONDITIONS, THIS ONE WAS BALE TO GIVE US BETTER RESULTS SHOWING THAT THE PREDCTION FOR THE CRASH TYPE IS VERY ACCURATE AND HIGH AT 96%

THE MODEL THAT PRODUCED CONSISTENT RESULTS WAS THE CONFUSION MATRIX AND THE DECISION TREE, I PRESUME LOGISTIC REGRESSIOON WASN'T BEST FIT BECAUSE IT IS BEST SUITED FOR BINARY CATEWGORIES

BETWEEN THE CONFUSION MATRIX GAVE A MORE ACCURATE DEPICTION OF THE RELATIONSHIP BETWEEN THE TARGET AND FEATURE VARIABLES

TO CHICAGO POLICE DEPARTMENT

Most of the accidents that the city of chicago experiences do not lead to fatal accidents, actually most of them the citizens are ok or with minor injuries

The intensity of injury is greatly affected by a lot of factors like

Timerelated (sunday and monday tend to have the least accidents)

Accidents occurring during the day are more likely to involve rear-end collisions

and pedestrians, while nighttime accidents have a similar distribution of fatalities.

Hit-and-run scenarios are more prevalent in fatal crashes

Weather related- most crashes happen when the weather is clear and rainy

Location based - most accidents happen on cicero drive having the highest count

Cause - most are identified to undetermined which closely related to the unharmed passangers in the injury category

Season- there is an identifiable crash on the 4th of july maybe infuence by 4th of july celebration

- **LIMITATIONS**

- ABSENCE OF CROSS VALIDATION AND REGULARIZING CLASS IMBALANCE
- ABSENCE OF FEATURE ENGINEERING THOUGH THE DATA HELD GRANULAR DATA
- NOT BINNING MY VARIABLES FOR FURTHER ANALYSIS

- **CONCLUSION**
- Through our detailed analysis and modeling of car crash data in Chicago, we've uncovered important insights into what contributes to accidents. By acting on these findings with targeted interventions, we can significantly cut down on the number and severity of car crashes. This means safer roads, fewer injuries, and a reduction in the financial burden that accidents place on our community. Ultimately, these steps will make Chicago a safer place for everyone.