PREDICTING THE NUMBER OF TRAFFIC ACCIDENT INJURIES IN CHICAGO[1]

Business Understanding.

Chicago is a busy city packed with traffic. All of this traffic dramatically raises the risk of a major car accident on any given roadway. On November 19, 2019, an unlicensed and uninsured driver failed to slow down at the intersection of 69th Street and Stony Island Avenue in Chicago, hitting and killing a bicyclist. On January 23, 2020, four people were killed in a crash involving a school bus, taxi, and a pickup truck at the intersection of Stony Island Avenue, 79th Street and South Chicago Avenue. The pickup driver ran a red light, first hit the taxi driver and then the bus driver, who was at the intersection. From 2019 to 2020, Chicago's traffic fatalities increased by 16 percent – making 2020 the deadliest year for Illinois drivers in 13 years. The high volume of traffic accidents, injuries, and deaths might lead one to question: what are the main causes of car accidents in Chicago?

The goal of this project is to produce a predictive model that would facilitate the analysis of the primary causes for car crashes in Chicago City. Using data provided by the City of Chicago, the aim is to use the information provided from the Chicago Data Portal and identify patterns and trends that can be implemented with satisfactory results. The goal is to reduce the instances of crashes with preventive measures informed by the predictive model.

Research Question

To produce a predictive model that would facilitate the analysis of the primary causes for car crashes in Chicago City to help reduce the instances of crashes with preventive measures informed by the predictive model.

Objectives

Main Objective

- To produce a predictive model that would facilitate the analysis of the primary causes for car crashes in Chicago City.

Sub Objectives

- Analyze control failures to identify opportunities for improvement
- Check for trends in the time of crash to relocate resources appropriately
- Check for accidents caused by poor road quality and conditions and how to improve on it.

Data Understanding.

Data Source.

The Traffic Crashes data comes from the Chicago Data Portal, an open data source maintained by the city of Chicago. The dataset contains all traffic crashes that were reported by the police within the city limits, going back to 2017. Linked to the

crash dataset are two datasets corresponding to Vehicles and Persons involved in the crash. Each crash incident has a unique crash record ID and report number associated with it, which allows for cross-referencing on the dashboards provided for the datasets. A link to the main dataset can be found here: https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if

Data Description.

We had three datasets which were:-

- Chicago car crashes (https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if)
- Vehicle data (https://data.cityofchicago.org/Transportation/Traffic-Crashes-Vehicles/68nd-jvt3)
- Driver / Passenger data (https://data.cityofchicago.org/Transportation/Traffic-Crashes-People/u6pd-qa9d).

Below are the descriptions of each column within the dataset.

1). Chicago Car Crash

Column Name	Description
crash_record_id	**Can be used to link to the same crash in the Vehicles and People datasets.**
rd_no	**Chicago Police Department report number**

crash_date	**Date and time of crash as entered by the reporting officer**
posted_speed_limit	**Posted speed limit, as determined by reporting officer**
Traffic_control_device	**Traffic control device present at crash location, as determined by reporting officer (signals, stop sign, etc)**
device_condition	**Condition of traffic control device, as determined by reporting officer**
weather_condition	**Weather condition at time of crash, as determined by reporting officer**
lighting_condition	**Light condition at time of crash, as determined by reporting officer**
first_crash_type	**Type of first collision in crash**
trafficway_type	**Trafficway type, as determined by reporting officer**
lane_ct	**Total number of through lanes in either direction, excluding turn lanes, as determined by reporting officer (0 = intersection)**
alignment	**Street alignment at crash location, as determined by reporting officer**
roadway_surface_cond	**Road surface condition, as determined by reporting officer**

road_defect	**Road defects, as determined by reporting officer**
crash_type	**A general severity classification for the crash. Can be either Injury and/or Tow Due to Crash or No Injury / Drive Away**
damage	**A field observation of estimated damage.**
prim_contributory_cause	**The factor which was most significant in causing the crash, as determined by officer judgment**
sec_contributory_cause	**The factor which was second most significant in causing the crash, as determined by officer judgment**
street_name	**Street address name of crash location, as determined by reporting officer**
num_units	**Number of units involved in the crash. A unit can be a motor vehicle, a pedestrian, a bicyclist, or another non-passenger roadway user. Each unit represents a mode of traffic with an independent trajectory.**
most_severe_injury	**Most severe injury sustained by any person involved in the crash**
injuries_total	**Total persons sustaining fatal, incapacitating, non-incapacitating, and possible injuries as determined by the reporting officer**

injuries_fatal	**Total persons sustaining fatal injuries in the crash**
injuries_incapacitating	**Total persons sustaining incapacitating/serious injuries in the crash as determined by the reporting officer. Any injury other than fatal injury, which prevents the injured person from walking, driving, or normally continuing the activities they were capable of performing before the injury occurred. Includes severe lacerations, broken limbs, skull or chest injuries, and abdominal injuries.**
injuries_non_incapacitating	**Total persons sustaining non-incapacitating injuries in the crash as determined by the reporting officer. Any injury, other than fatal or ncapacitating injury, which is evident to observers at the scene of the crash. Includes lump on head, abrasions, bruises, and minor lacerations.**
crash_hour	**The hour of the day component of CRASH_DATE.**
crash_day_of_week	**The day of the week component of CRASH_DATE. Sunday=1**
latitude	**The latitude of the crash location, as determined by reporting officer, as derived from the reported address of crash**

longitude	**The longitude of the crash location, as determined by reporting officer, as derived from the reported address of crash**

2) Driver/Passenger data

Column Name	Description
crash_record_id	**This number can be used to link to the same crash in the Crashes and Vehicles datasets. This number also serves as a unique ID in the Crashes dataset.**
person_type	**Type of roadway user involved in crash**
rd_no	**Chicago Police Department report number. For privacy reasons, this column is blank for recent crashes.**
crash_date	**Date and time of crash as entered by the reporting officer**
seat_no	**Code for seating position of motor vehicle occupant: 1= driver, 2= center front, 3 = front passenger, 4 = second row left, 5 = second row center, 6 = second row right, 7 = enclosed passengers, 8 = exposed passengers, 9= unknown position, 10 = third row left,**

	11 = third row center, 12 = third row right
city	**City of residence of person involved in crash**
state	**State of residence of person involved in crash**
zipcode	**ZIP Code of residence of person involved in crash**
sex	**Gender of person involved in crash, as determined by reporting officer**
age	**Age of person involved in crash**
drivers_license_state	**State issuing driver's license of person involved in crash**
drivers_license_class	**Class of driver's license of person involved in crash**
safety_equipment	**Safety equipment used by vehicle occupant in crash, if any**
airbag_deployed	**Whether vehicle occupant airbag deployed as result of crash**
ejection	**Whether vehicle occupant was ejected or extricated from the vehicle as a result of crash**
injury_classification	**Severity of injury person sustained in the crash**

Driver_action	**Driver action that contributed to the crash, as determined by reporting officer**
driver_vision	**What, if any, objects obscured the driver's vision at time of crash**
physical_condition	**Driver's apparent physical condition at time of crash, as observed by the reporting officer**
pedpedal_action	**Action of pedestrian or cyclist at the time of crash**
pedpedal_visibility	**Visibility of pedestrian or cyclist safety equipment in use at time of crash**
pedpedal_location	**Location of pedestrian or cyclist at the time of crash**
bac_result	**Status of blood alcohol concentration testing for driver or other person involved in crash**
bac_result value	**Driver's blood alcohol concentration test result (fatal crashes may include pedestrian or cyclist results)**
cell_phone_use	**Whether person was/was not using cellphone at the time of the crash, as determined by the reporting officer**

3) Vehicle data

Column Name	Description
crash_record_id	This number can be used to link to the same crash in the Crashes and People datasets. This number also serves as a unique ID in the Crashes dataset.
rd_no	Chicago Police Department report number. For privacy reasons, this column is blank for recent crashes.
crash_date	Date and time of crash as entered by the reporting officer
unit_type	The type of unit (i.e Driver, parked, pedestrian, bicycle, etc)
num_passengers	Number of passengers in the vehicle. The driver is not included. More information on passengers is in the People dataset.
make	The make (brand) of the vehicle, if relevant
model	The model of the vehicle, if relevant
lic_plate_state	The state issuing the license plate of the vehicle, if relevant
vehicle_year	The model year of the vehicle, if relevant
vehicle_defect	Indicates part of car containing defect (brakes, wheels, etc.)

vehicle_type	**The type of vehicle, if relevant (passenger, truck, bus, etc)**
vehicle_use	**The normal use of the vehicle, if relevant**
maneuver	**The action the unit was taking prior to the crash, as determined by the reporting officer**
towed_I	**Indicator of whether the vehicle was towed**
occupant_cnt	**The number of people in the unit, as determined by the reporting officer**
exceed_speed_limit_I	**Indicator of whether the unit was speeding, as determined by the reporting officer**
First_contact_point	**Indicates orientation on car that was hit (front, rear, etc)**

Data Preparation

- **Loading the data**

Lets use the .read_csv() format to read our datadets.

```
# read the crashes data
crashes_df = pd.read_csv("Traffic_Crashes_-_Crashes.csv")
```

```
# read the people data
people_df = pd.read_csv("Traffic_Crashes_-_People.csv")
```

```
C:\Users\USER\anaconda3\envs\learn-env\lib\site-packages\IPython
(29) have mixed types.Specify dtype option on import or set low_
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name
```

```
# read the vehicles data
vehicles_df = pd.read_csv("Traffic_Crashes_-_Vehicles.csv")
```

```
C:\Users\USER\anaconda3\envs\learn-env\lib\site-packages\IPython
(19,21,40,41,42,44,48,49,50,53,55,58,59,61,71) have mixed types.
  has_raised = await self.run_ast_nodes(code_ast.body, cell_name
```

- **Cleaning the data**

1. ***Checked information of each dataset.***
2. ***Dropped un-necessary columns on each dataset.***
3. ***Checked new information of each dataset after dropping the columns.***
4. ***Checked the shape of each dataset.***
5. ***Checked for missing values of each dataset in which some columns and missing data were dropped.***
6. ***Checked for unique categories in our three datasets.***
7. ***Merged our three datasets.***
8. ***Cleaned our merged dataset by dropping unnecessary columns and missing values***
9. ***Checked for unique categories on our merged dataset.***
10. ***Did feature engineering to our target variable and predictors.***

11. ***Explored our data by graphing different factors and explaining the output.***
12. ***Did correlation.***
13. ***Did feature importance using Random Forest to answer the project's objectives.***
14. ***Encoded the categorial columns to numerical values.***
15. ***Split our data before modeling the data.***

Modeling.

Multiclass classification models were used in order to try to classify these incidents. Models such as Logistic regression, K-Nearest Neighbors and Decision Trees were among those used. All of the models had difficulty discerning between improper driving and external factors, which was rightfully so. There were a number of incident causes that would not be picked up by the data columns that were provided.

1). Linear Regression

The baseline_log_loss is **0.5833135603648449.**

The accuracy is **0.6613448648648649**

The recall is **0.2526898247771288**

The precision is **0.49695198750566777**

The AUC is **0.7097859344407482**

In this model our accuracy was **66%** and AUC is **70%**

2). K Nearest Neighbour

The baseline_log_loss is **1.7893844751451298.**

The accuracy is **0.7402032432432433.**

The recall is **0.5526180961164053.**

The precision is **0.6317392450288458.**

The AUC is **0.6521851756929807**

In this model our accuracy was **74%** and AUC is **65%**

3). Decision Tree

The baseline_log_loss is **6.625605348564119.**

The accuracy is **0.8410118918918918.**

The recall is **0.7712880418075623.**

The precision is **0.7610140788110102.**

The AUC is **0.6973719404360946**

In this model our accuracy was **84%%** and AUC is **70%**

Evaluation

As for the final model, DecisionTreeClassifier was the best fit model. This is because:-

Accuracy score is **84%** which is the highest as compared to KNN and Logistic Regression.

Precision score is **76%** which is the highest as compared to KNN and Logistic Regression.

Recall score is **77%** which is the highest as compared to KNN and Logistic Regression.

Can conclude that the higher the percentage, the better the model is performing as one can verify from the accuracy, precision and recall score.

Conclusions

- Road traffic injury is a threat to health and development.
- No institutional frameworks exist within the region.
- Legislation and enforcement of key road safety interventions need to be strengthened in many regions.
- Adherence to vehicle and road design safety standards is low in the region.
- Some important data are non-existent or incomplete.

Recommendations

- Establish and strengthen lead agencies and manage performance through target setting.
- Make safe, healthy, environment-friendly transport choices; design transport around walking, cycling and public transport.
- Focus on implementing the five most effective interventions to reduce chances of injury during a crash.

* **Control speed.** Speed limits on urban and rural roads and motorways should be set by defining each road type in the regions.

* **Implement seat belt laws.** Seat belt laws should apply to all vehicle occupants; and these laws should be better enforced.

* **Enforce use of standard motorcycle helmets.** Motorcycle helmet law should include pillion riders and a standard needs to be defined for these helmets in the regions.

* Ascertain the role of alcohol in road crashes and control it, if found to be a problem.

- Allow only safe vehicles on the roads. Vehicle manufacturing and import standards should be evaluated in the region to ensure that only vehicles that allow international safety standards.

- Ensure safe road design through safety audits at all stages of road construction and maintenance.
- Improve trauma care. A pre-hospital care system of ambulances connected through a universal access number is an important but just one component of a comprehensive trauma system.
- Define data needs; harmonize definitions and data collection methodology in order to advocate the need for road safety targets.
- Enhance institutional capacity for data gathering, analysis and dissemination.
- Implementation of collaborative relationships between health, police and traffic authorities will need to be established for setting up surveillance systems.