C 2

SUPERVISED LEARNING

SeiedeShiva Moosavi

Agenda

- Introduction
- Data cleaning
- Features
- Data exploration
- Feature selections
- Machine learning Models
- Comparisons
- Conclusions
- Future work
- Reference

Introduction

In the research, we want to do some data exploration to answer a question by using machine learning models.

The data source is Kaggle and the format is SQL. This data comes from the California Highway Patrol and covers collisions from January 1st, 2001 until mid October 2020.

Introduction

There are three main tables:

- Collisions: Contains information about the collision, where it happened, what vehicles were involved.
- Parties: Contains information about the groups people involved in the collision including age, sex, and sobriety.
- Victims: Contains information about the injuries of specific people involved in the collision.

In this project, the three main tables joined where collision date is after 2020-01-01

Data Cleaning

- NAN values: Columns which have more than 50% unknown value have been dropped.
- Outliers: Rows with parties' age of zero have been dropped (32506 out of 351975)

Features

There are 86 columns including categorical and numeric features such Party Number, Victim Sex, Victim Age, Party Race, Vehicle Year, Vehicle Make, Party Age, Party Sex, At Fault

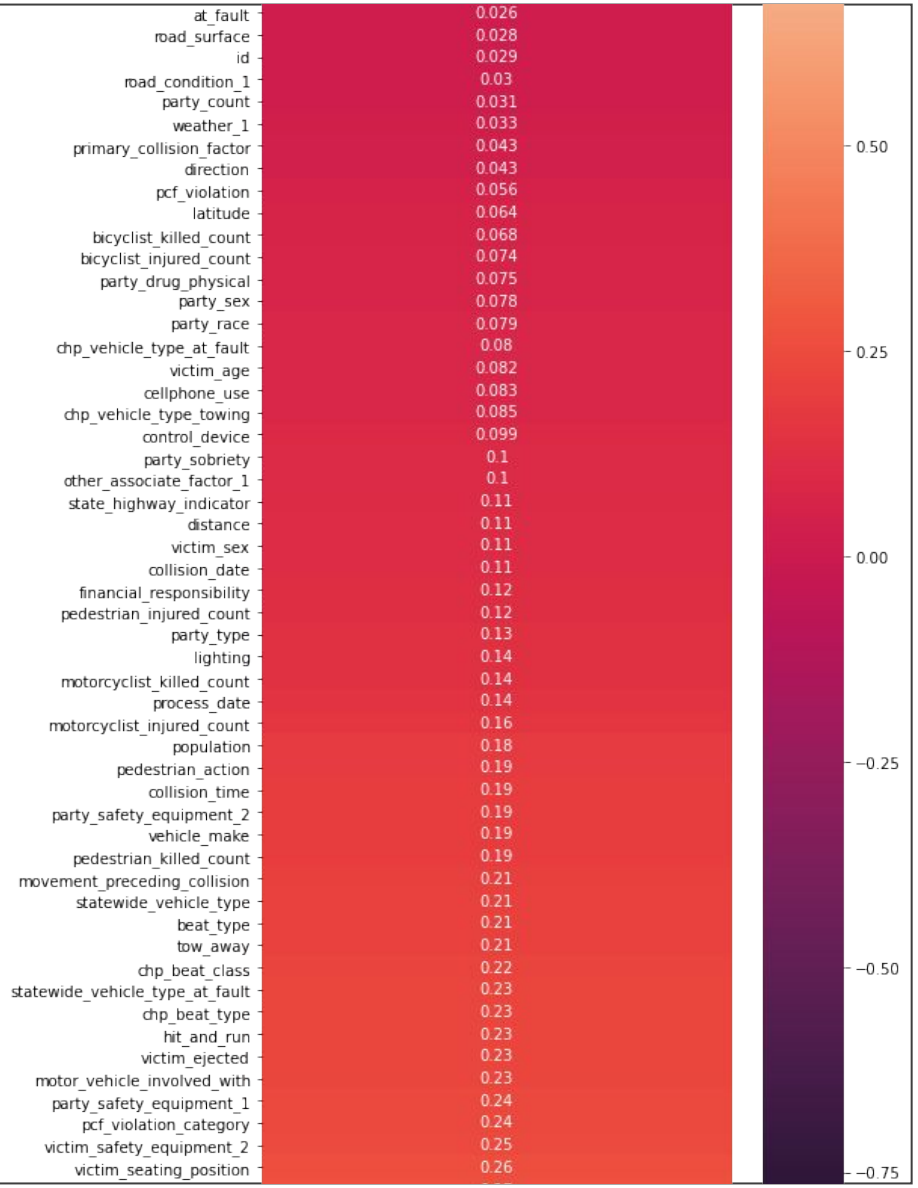
party_sobriety	A	B	C	D	G	H
party_drug_physical						
E	89	473	14	4	0	0
F	179	7	7	2	0	0
G	0	0	0	0	21596	0
H	0	0	0	0	0	18630
I	1144	56	39	2	0	0
NA	278607	14686	3087	1957	0	0

Data exploration

After considering the table of parties' use of drugs and the parties' sobriety table, is has been notified that most the related unknown on parties' use of drugs are related to sober parties

FEATURE SELECTIONS

Columns with correlation more than 0.25 has been chosen.



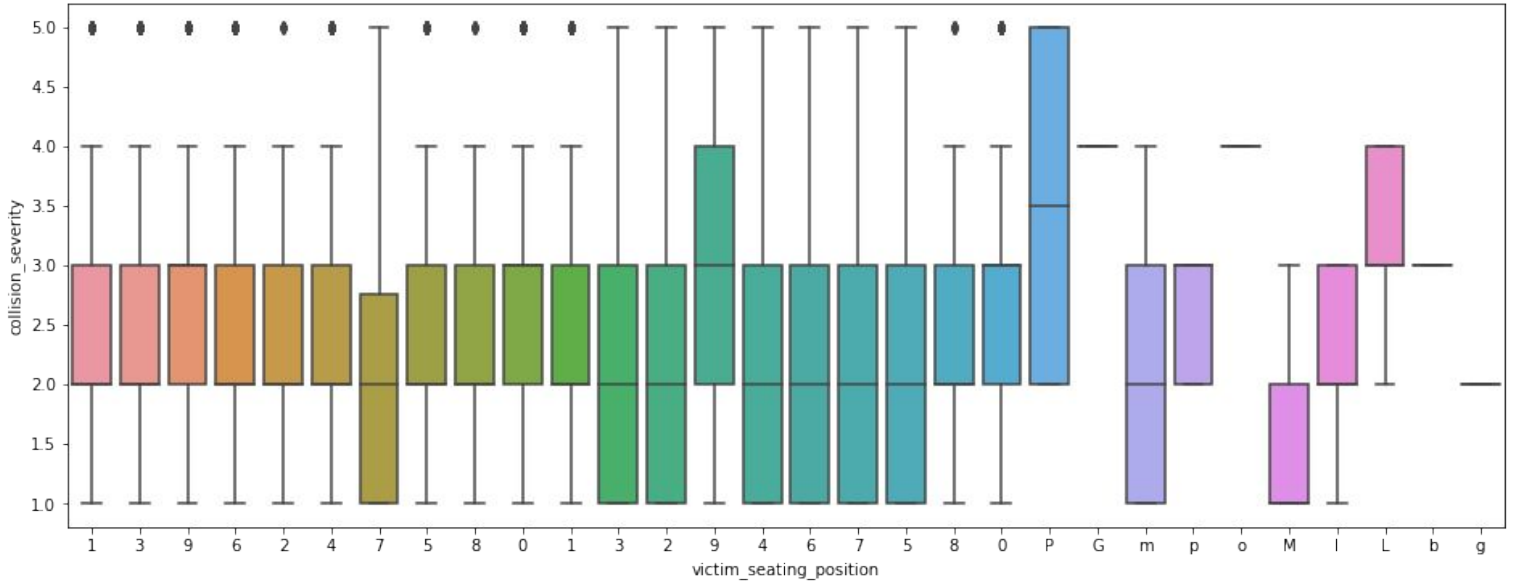
Feature selections

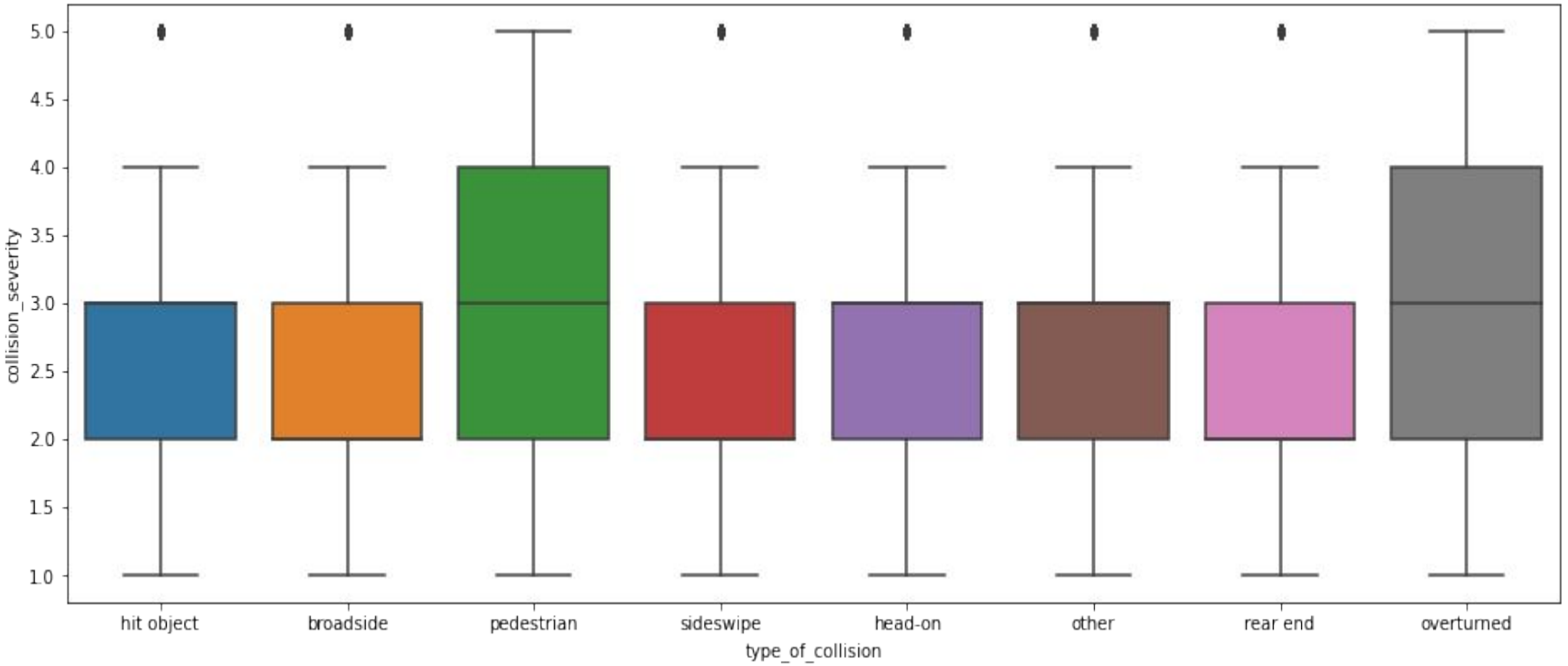
Columns with correlation more than 0.25 has been chosen.

Victim seating position, type of collision, victim role, victim safety equipment 1, primary road, secondary road columns have been chosen to predict their effects on collision severity

Feature selections Victim seating position

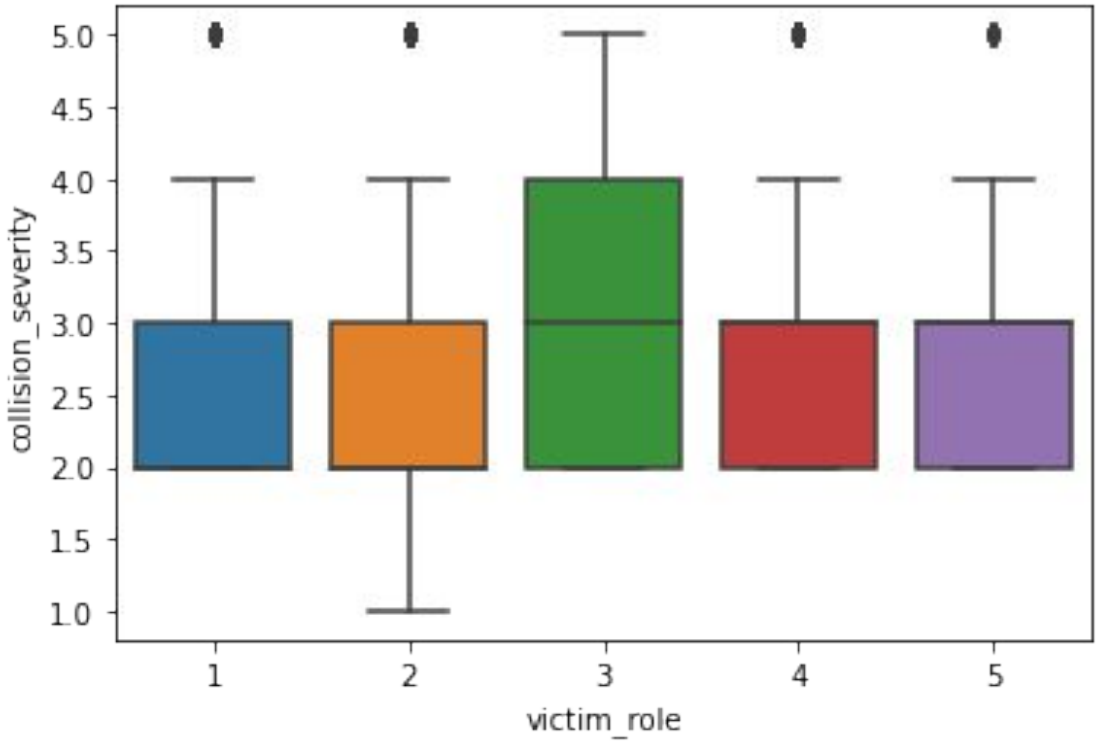
- 1 - Driver
- 2 thru 6 - Passengers
- 7 - Station Wagon Rear
- 8 - Rear Occupant of Truck or Van
- 9 - Position Unknown
- 0 - Other Occupants
- A thru Z - Bus Occupants
- - - Not Stated





Feature selections type of collision

- A - Head-On
- B - Sideswipe
- C - Rear End
- D - Broadside
- E - Hit Object
- F - Overturned
- G - Vehicle/Pedestrian
- H - Other
- - - Not Stated

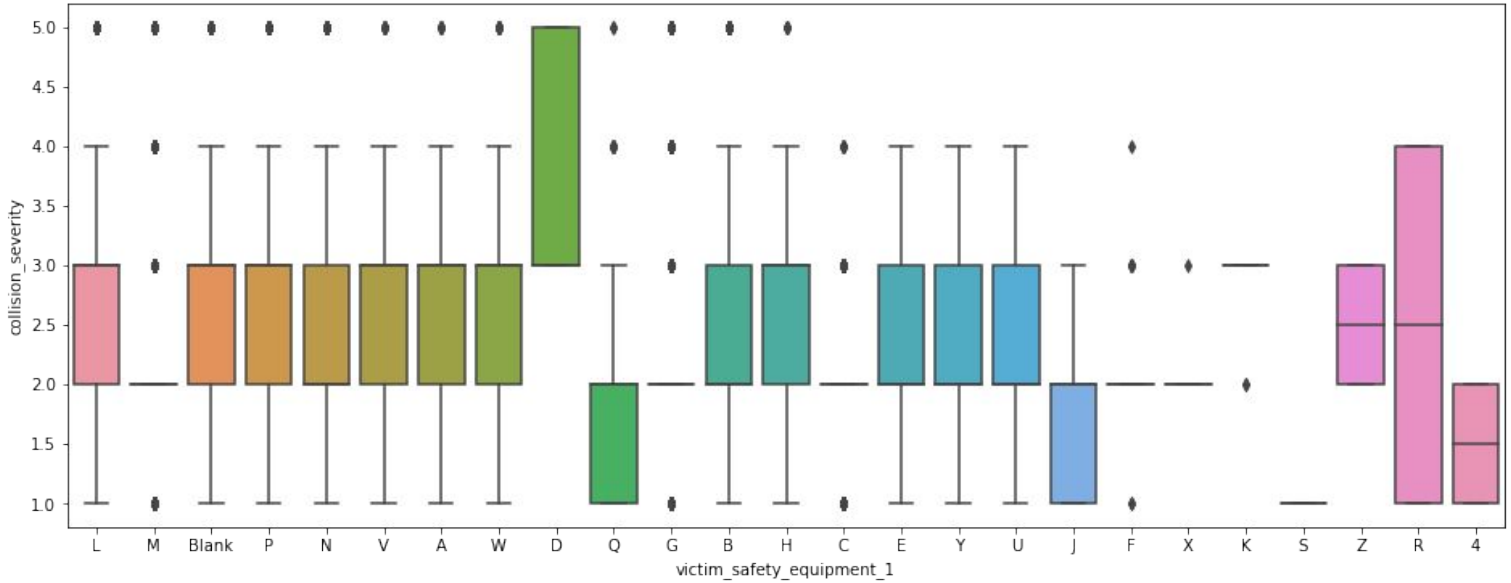


Feature selections victim role

- 1 - Driver
- 2 - Passenger (includes non-operator on bicycle or any victim on/in parked vehicle or multiple victims on/in non-motor vehicle)
- 3 - Pedestrian
- 4 - Bicyclist
- 5 - Other (single victim on/in non-motor vehicle; e.g. ridden animal, horse-drawn carriage, train, or building)
- 6 - Non-Injured Party

- A - None in Vehicle
- B - Unknown
- C - Lap Belt Used
- D - Lap Belt Not Used
- E - Shoulder Harness Used
- F - Shoulder Harness Not Used
- G - Lap/Shoulder Harness Used
- H - Lap/Shoulder Harness Not Used
- J - Passive Restraint Used
- K - Passive Restraint Not Used
- L - Air Bag Deployed
- M - Air Bag Not Deployed
- N - Other
- P - Not Required
- Q - Child Restraint in Vehicle Used
- R - Child Restraint in Vehicle Not Used
- S - Child Restraint in Vehicle, Use Unknown
- T - Child Restraint in Vehicle, Improper Use
- U - No Child Restraint in Vehicle
- V - Driver, Motorcycle Helmet Not Used
- W - Driver, Motorcycle Helmet Used
- X - Passenger, Motorcycle Helmet Not Used
- Y - Passenger, Motorcycle Helmet Used
- - or blank - Not Stated

Feature selections victim safety equipment1



'QUAIL CREEK RD', 'HARRIS RD', 'SIERRA AVE', ..., 'VANGUARD DR', 'EMERSON RD', 'GATESIDE CT'

Feature selections primary road

'NORRIS RD', 'GOSFORD RD', 'FIRST ST',

..., 'LOS ABITOS BL', 'WESTAR DR',

'VENTURA RD FRONTAGE RD'

The number of collisions happened are more than 100 times.

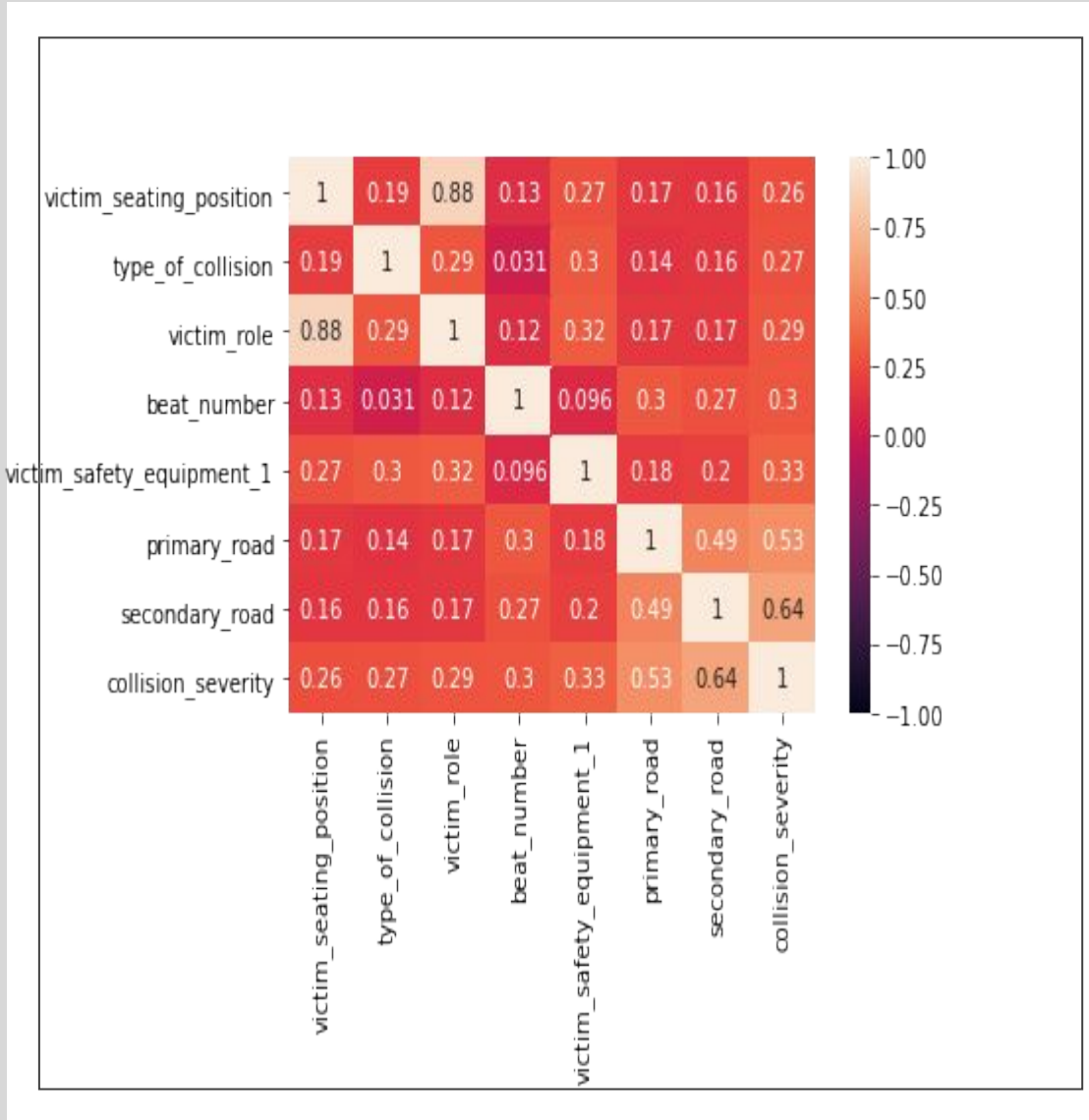
primary_road	avg_severity	count
INTERSTATE 5 SOUTHBOUND	3.421965	173
SR-39 (SAN GABRIEL CANYON RD.)	3.398230	113
US-395	3.382979	141
SR-4	3.329341	167
GENEVA AV	3.304348	161
...	...	...
INTERNATIONAL BL	1.780769	260
WALNUT AV	1.731544	149
WILSON WY	1.710744	121
MARCH LN	1.699605	253
SIERRA AV	1.630695	417

Secondary road	Avg severity	count
FRANCISQUITO AVE	4.379747	158
LAKEWOOD BOULEVARD	3.227941	136
ROXFORD ST	2.950820	183
VAN NESS AV	2.926966	178
TYLER ST	2.898734	158
...	...	...
OLIVE AV	1.829268	123
BANCROFT AV	1.771930	114
CITRUS AV	1.767123	146
10TH ST WEST	1.729508	122
SIERRA AV	1.526946	167

Feature selections secondary road

`'QUAIL CREEK RD', 'HARRIS RD', 'SIERRA AVE', ..., 'VANGUARD DR', 'EMERSON RD', 'GATESIDE CT'`

The number of collisions happened are more than 100 times.



Correlation

Columns with correlation more than 0.25

Machine learning Models

- Random Forest
- Logistic Regression

RANDOM FOREST

TUNING

The accuracy of the predictor on the training and test data.

Accuracy= 96%

```
Train score: 0.9646983585589427
Test score: 0.9646704374271665
```

RANDOM FOREST

IMPORTANCE FOR INTERPRETABILITY

feat	importance
Victim role	0.449560
Victim safety equipment_1	0.241410
Victim seating position	0.143473
Secondary road	0.115767
Primary road	0.034933
Type of collision	0.014857

RANDOM FOREST CONFUSION MATRIX

Pred	property damage only	pain	other injury	severe injury	fatal
Actual					
property damage only	6819	28	29	10	0
pain	0	36178	25	16	0
other injury	0	665	19928	14	0
severe injury	63	352	545	4493	0
fatal	66	74	194	405	462

Logistic Regression

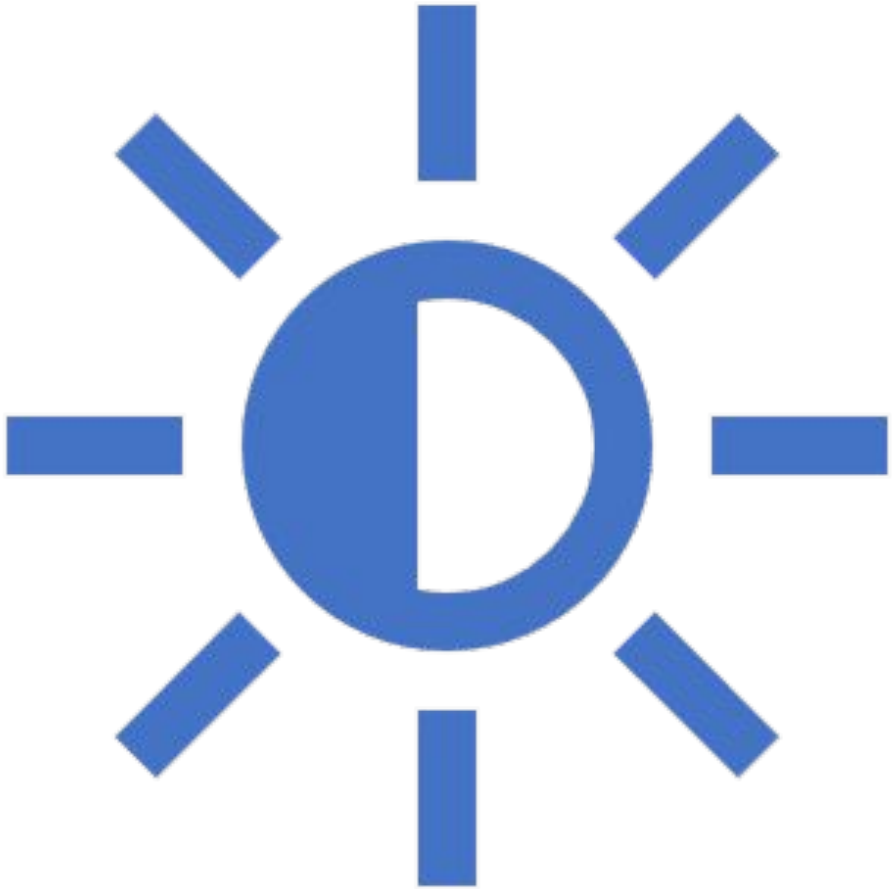
MODEL BEST PARAMS:
C = 1000

MODEL.SCORE(X_TRAIN, Y_TRAIN)=
0.6853975698145385

MODEL.SCORE(X_TEST, Y_TEST)=
0.6872637353267202

LOGISTIC REGRESSION CONFUSION MATRIX

Pred	property damage only	pain	other injury	severe injury	fatal
Actual					
property damage only	3701	3013	170	2	0
pain	815	31109	4189	102	4
other injury	56	8668	11549	327	7
severe injury	3	1026	2703	1668	53
fatal	0	106	364	398	333



Comparisons

Random Forest Model can predict the outcome better than Logistic Regression model.

Conclusions

The outcome of the research shows that Victim seating position, type of collision, victim role, victim safety equipment 1, primary road, secondary road are important factors and the most important is victim role on severity of collision.

In the research, a model to predict collision severity based on important features are represented.

The results are important for DMV, Police, or even regular people to manage reducing severity of collisions.

Future work

Our model can be improved by using more features, more data exploration, and more machine learning models, we can answer to some interesting and useful questions such as:

- When do different makes and models of motorcycles crash?
- On what days are pedestrians involved in collisions?
- Are DUIs more likely on certain days?
- How has COVID19 changed collisions? Are different types of vehicles involved? At different times?
- What type/color of car/motorcycle is involved in the most crashes?
- Can we predict injuries based on the other variables in a collision?
- Are their more accidents around sunrise/sunset?

Reference

California Traffic Collision Data from SWITRS | Kaggle