IBM

Coursera Capstone project

Lokesh mulpuru

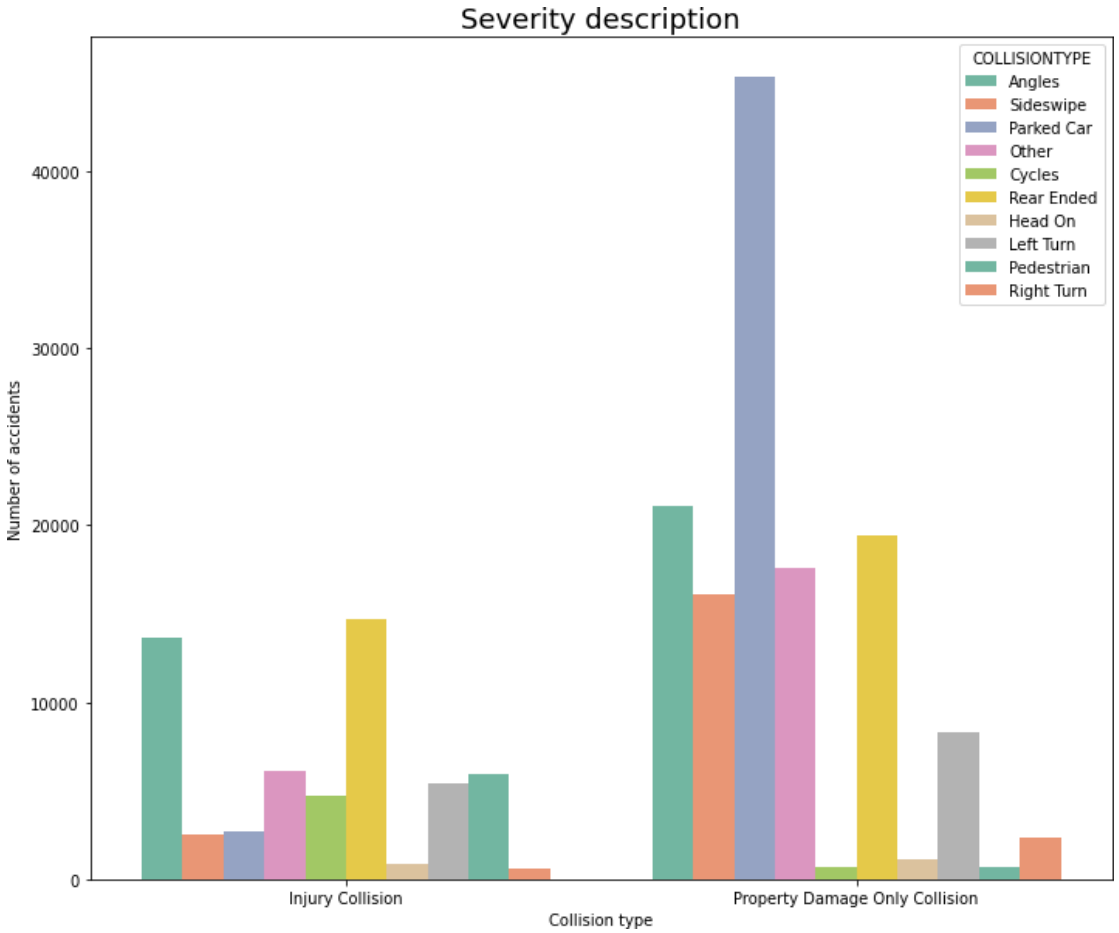
Data understanding

- Data source:

  https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

- **No. of attributes : 38**
- **No. of entries: 194673**
- **Severity of accidents – Injury / Property damage**
- Location types: **Alley / Block / Intersection**
- Human factors: **Inattentiveness, under-influence, speeding**
- Location factors: **Weather, road and light conditions**

Severity of accidents and collision type

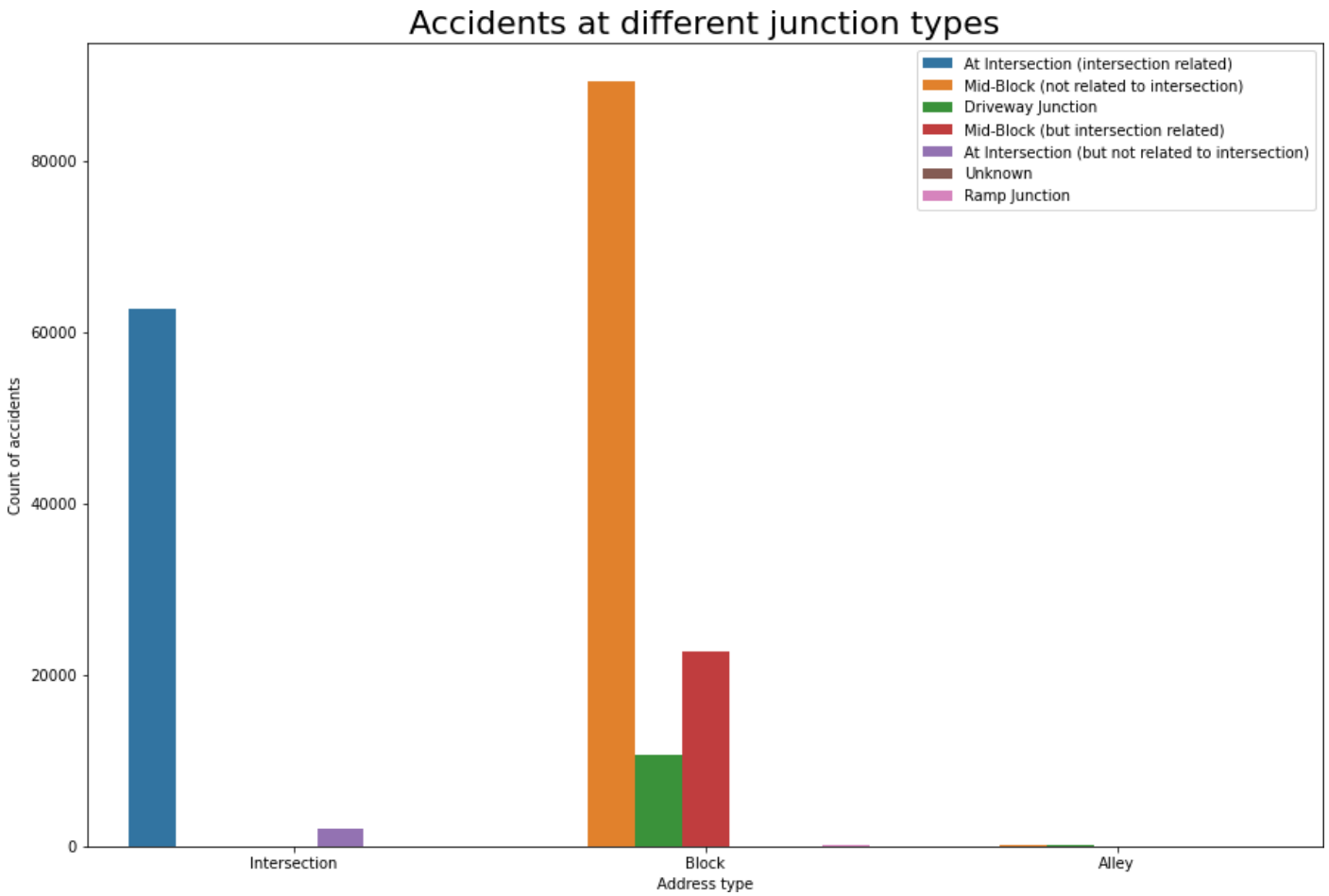


1. More accidents occur on weekdays

2.In injury collision, major accidents occur due to vehicles hitting another vehicle's rear end or hitting pedestrians

3. In property damage collisions, mostly parked cars are hit

Where do most accidents occur



1. Some accidents unrelated to intersections occur at intersections.
2. In blocks, maximum accidents occur at midblock and away from intersections.
3. Very few accidents occur at alleys

Correlation between severity of accident and location

	SEVERITYCODE	Intersection	Alley	Block
SEVERITYCODE	1.000	0.199	-0.026	-0.185
Intersection	0.199	1.000	-0.044	-0.970
Alley	-0.026	-0.044	1.000	-0.085
Block	-0.185	-0.970	-0.085	1.000

Maximum road accidents occur at intersections

Correlation between severity of accident and people affected

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
SEVERITYCODE	1.000	0.131	0.246	0.214	-0.055
PERSONCOUNT	0.131	1.000	-0.023	-0.039	0.381
PEDCOUNT	0.246	-0.023	1.000	-0.017	-0.261
PEDCYLCOUNT	0.214	-0.039	-0.017	1.000	-0.254
VEHCOUNT	-0.055	0.381	-0.261	-0.254	1.000

Pedestrians and pedestrians on cycle are most affected due to road accidents

Different machine learning models' comparison for regression

Model	Accuracy score(%)	F1 score(%)	Precision(%)	Recall(%)
Logistic Regression	74.99	36.48	76.67	23.94
Decision Tree	75.24	37.41	77.43	24.67
Random Forest	75.21	37.31	77.27	24.59
SVM	74.84	30.62	88.68	18.51
XGBoost	75.26	35.56	81.35	22.75

Discussion

- Maximum accidents occur during weekdays at intersections
- Weather conditions do not play a significant role in accidents
- Road and lighting conditions have a weak correlation with accidents
- Being under influence doesn't cause noticeably more accidents than being inattentive
- Between blocks, maximum accidents occur at mid-blocks
- In collision accidents, maximum damage is done to parked cars

In this project I have identified the relation between accidents and several human, environmental and location attributes. Maximum accidents occur at intersections related to pedestrians or cyclists. I analysed different machine learning models to classify accidents as injury or collision accidents. The " XGBoost" model offered maximum accuracy. It correctly predicted 81.35% as injury collisions. This data could be used by governments to establish separate signals for allowing pedestrians and cyclists to cross at intersections. It is also aimed at us, whether we are pedestrians, cyclists or vehicle owners to be more careful at intersections to prevent an accident.

Conclusions

I could achieve an accuracy of ~75% using the XGB Classifier. There are a lot of variances which have not been accounted for. However, using this project we could really narrow down to the location(intersections) where maximum accidents occur and the most affected. We also understood that there is very less importance of human and weather factors in causing an accident. The prediction could be improved by capturing real time data during accidents.