Predicting Car Accident Severity

Capstone Project

Benitez Samuel

August 26, 2020

1. *Introduction*

1.1. *Background*

According to the World Health Organization (WHO), every year, traffic accidents kill approximately 1.3 million people around the world and between 20 and 50 million suffer non-fatal injuries. It represents one of the main causes of death in all age groups, and the first among people between 15 and 29 years old.

Pedestrians, cyclists, and riders of motorized 2- and 3-wheelers account for half of all these deaths and are known as "vulnerable road users."

Therefore, road accidents represent a public health problem worldwide that must be treated in different ways depending on the road culture of the countries.

1.2. *Business Problem*

Today, having a vehicle is practically a necessity for people anywhere in the world. However, due to the large number of people, means of transport and different climatic or circumstantial factors, road accidents are part of the day to day in cities around the world. For this reason, it is important for any driver to be able to avoid an accident, especially those of high severity, which can result in a catastrophic end.

Being able to predict the severity or probability of an accident is highly useful for the well-being of any driver or pedestrian. This would undoubtedly help all drivers to have a greater perspective on possible accidents and thus make the best decision about whether to drive, not drive or do it safely.

1.3. *Interest*

The main objective is to identify the severity (1 or 2) or probability of a possible accident, so this work can be useful for anyone interested in assessing risks while driving, the traffic control departments, governments or simply to educate anyone who drives a vehicle.

2. *Data Acquisition and Cleaning*

2.1. *Data Source*

For the analysis and the creation of a predictive model using supervised machine learning, the Seattle Collisions Dataset given by the Seattle Department of Transportation (SDOT) with the reports of road accidents will be used.

2.2. *Data Description*

Accident details

- `SEVERITYCODE`: severity of accident, target variable to predict.
  - 1: *Property Damage Only*
  - 2: *Injury*
- `SEVERITYDESC`: complementary description to the code.
- `COLLISIONTYPE`: type of collision, general description.

Location and time

- `LOCATION`: general address.
- `X`, `Y`: exact location of the accident.
- `ADDRTYPE`: whether the accident occurred in an alley, block or intersection.
- `JUNCTIONTYPE`: category of junction, for example: mid-block, intersection, driveway, etc.
- `INCDATE`, `INCDTTM`: date and time of the incident.

People affected

- `PERSONCOUNT`: total number or people involved.

- PEDCOUNT, PEDROWNOTGRNT: number of pedestrians involved and whether their right of way was granted or not.
- PEDCYLCOUNT: number of bicycles involved.
- VEHCOUNT, HITPARKEDCAR: number of vehicles involved and whether it was a parked car or not.

Driver fault

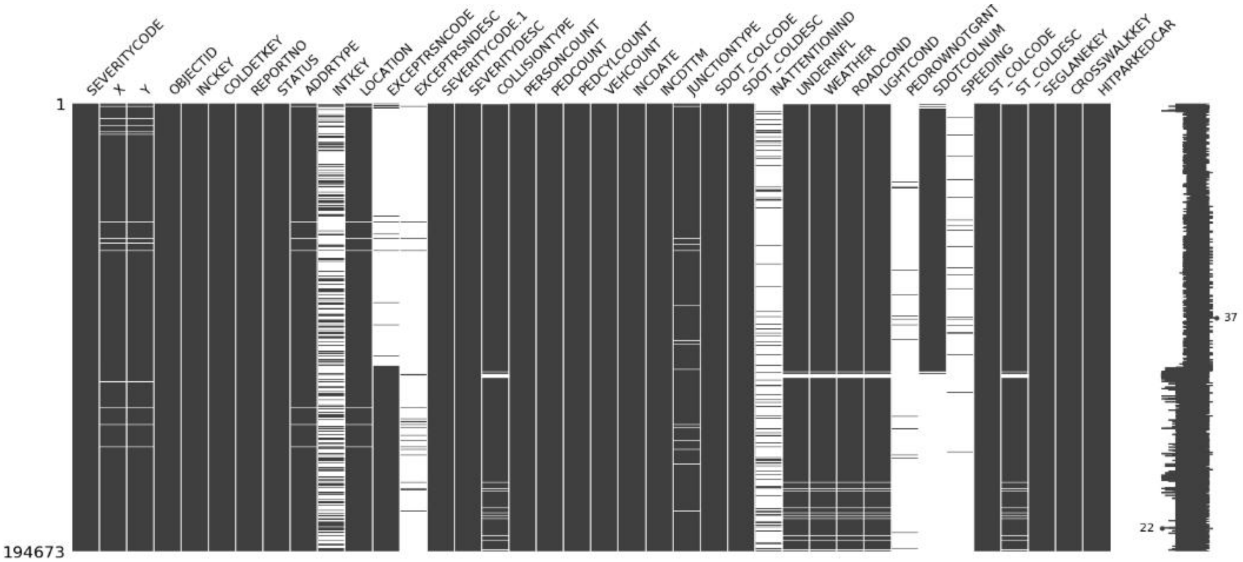
- INATTENTIONIND: whether the driver was inattentive.
- UNDERINFL: whether the driver was under influence of drugs or alcohol.
- SPEEDING: whether the drives was speeding.

Environmental issues

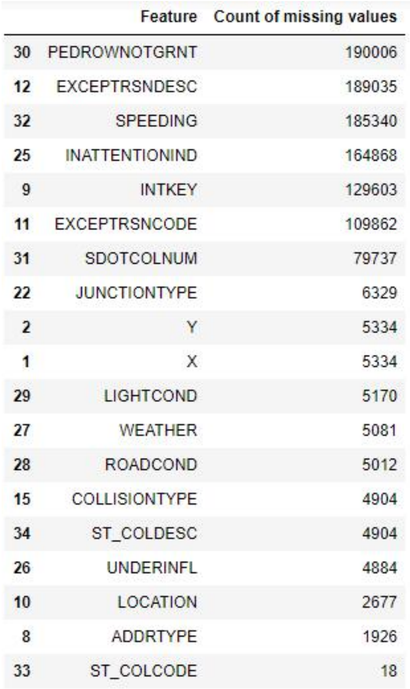
- WEATHER: general description of the weather condition during the accident: clear, raining, snowing, etc.
- ROADCOND, LIGHTCOND: the condition of the road and light, for example: dry or wet and daylight or dark.

2.3. Data Cleaning

Data before cleaning



Missing Data: exact values

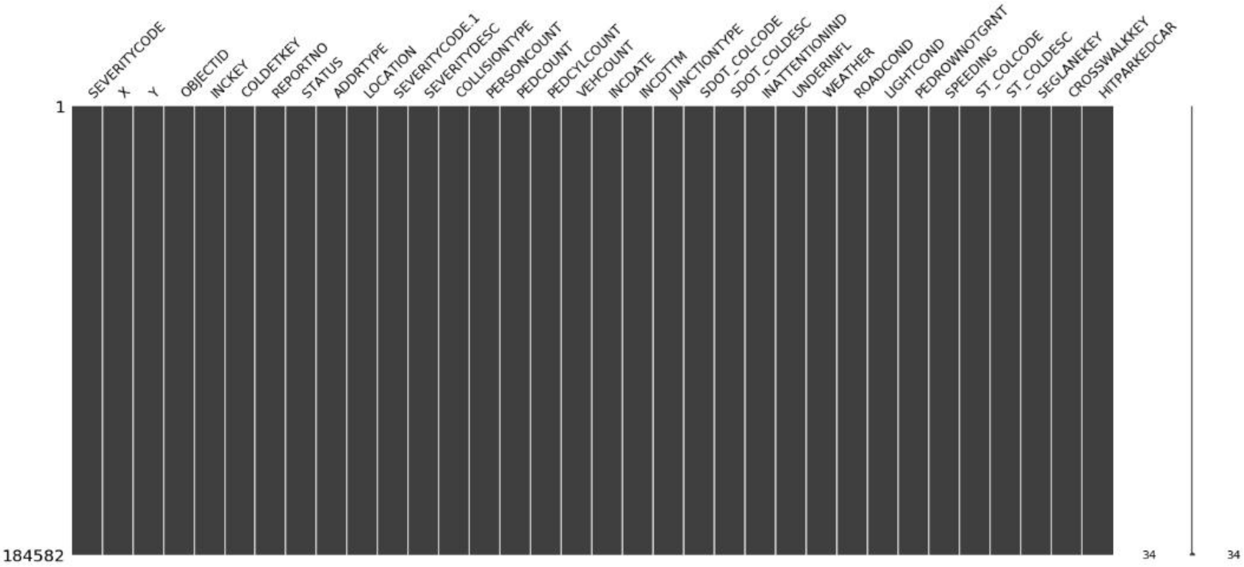


	Feature	Count of missing values
30	PEDROWNOTGRNT	190006
12	EXCEPTRSNDESC	189035
32	SPEEDING	185340
25	INATTENTIONIND	164868
9	INTKEY	129603
11	EXCEPTRSNCODE	109862
31	SDOTCOLNUM	79737
22	JUNCTIONTYPE	6329
2	Y	5334
1	X	5334
29	LIGHTCOND	5170
27	WEATHER	5081
28	ROADCOND	5012
15	COLLISIONTYPE	4904
34	ST_COLDESC	4904
26	UNDERINFL	4884
10	LOCATION	2677
8	ADDRTYPE	1926
33	ST_COLCODE	18

We filled in missing values with consistent data if possible and dropped rows and columns with important data missing. Also, we changed some variable types to be consistent, such as “UNDERINFL”, for example, attribute with 4 different values of different type.

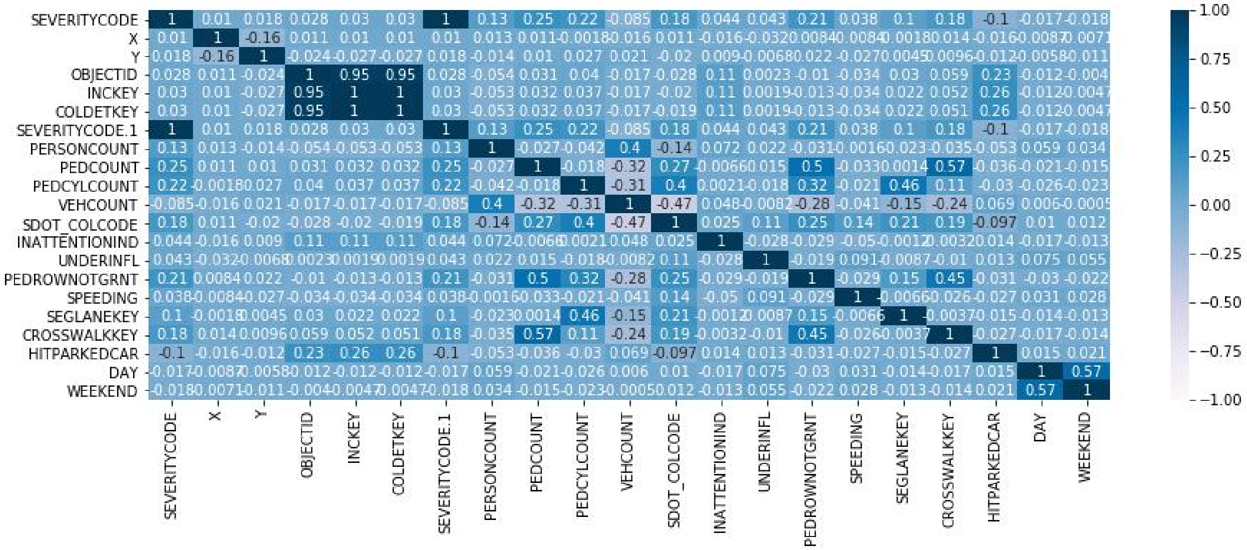
Then format every feature as *int64* (with binary values 1 or 0) if possible or type *str* for the categorical variables, which later will be converted to unique columns using one hot encoding.

Data after cleaning and formatting



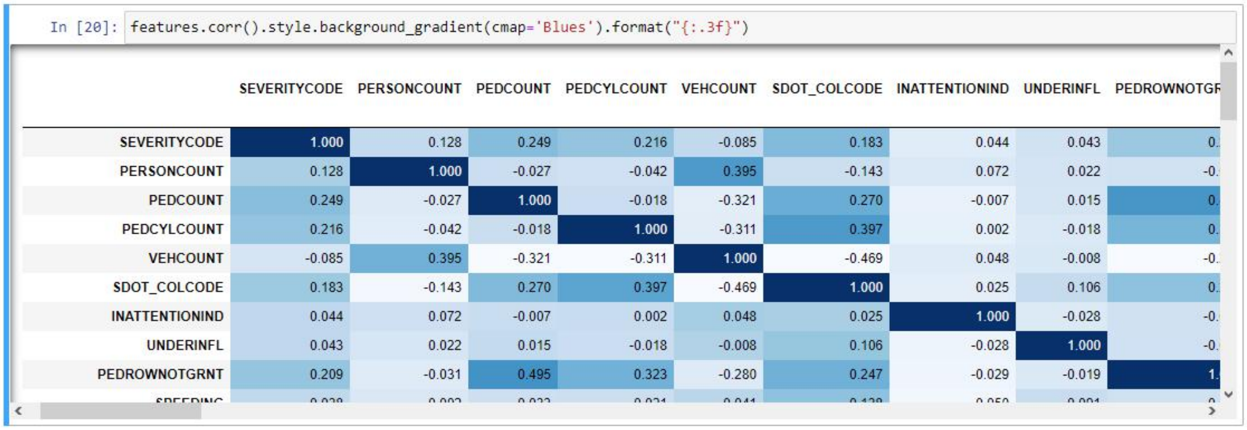
2.4. *Feature Selection*

After data cleaning, formatting and encoding, there were 184,582 rows and 60 features, from which we selected the first 10 most correlated with *SEVERITYCODE* for modeling.



3. *Exploratory Data Analysis*

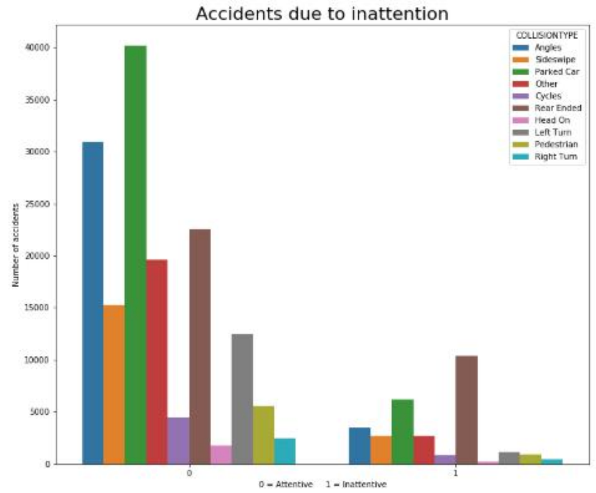
3.1. *Relationships between the target variable*

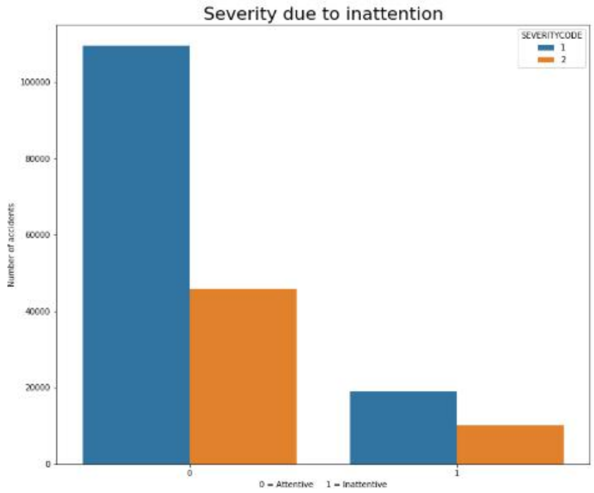


```
In [20]: features.corr().style.background_gradient(cmap='Blues').format("{:.3f}")
```

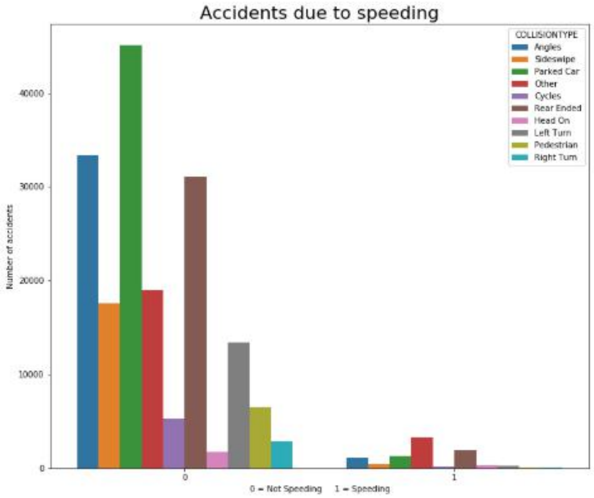
	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	SDOT_COLCODE	INATTENTIONIND	UNDERINFL	PEDROWNOTGR
SEVERITYCODE	1.000	0.128	0.249	0.216	-0.085	0.183	0.044	0.043	0.
PERSONCOUNT	0.128	1.000	-0.027	-0.042	0.395	-0.143	0.072	0.022	-0.
PEDCOUNT	0.249	-0.027	1.000	-0.018	-0.321	0.270	-0.007	0.015	0.
PEDCYLCOUNT	0.216	-0.042	-0.018	1.000	-0.311	0.397	0.002	-0.018	0.
VEHCOUNT	-0.085	0.395	-0.321	-0.311	1.000	-0.469	0.048	-0.008	-0.
SDOT_COLCODE	0.183	-0.143	0.270	0.397	-0.469	1.000	0.025	0.106	0.
INATTENTIONIND	0.044	0.072	-0.007	0.002	0.048	0.025	1.000	-0.028	-0.
UNDERINFL	0.043	0.022	0.015	-0.018	-0.008	0.106	-0.028	1.000	-0.
PEDROWNOTGRNT	0.209	-0.031	0.495	0.323	-0.280	0.247	-0.029	-0.019	1.

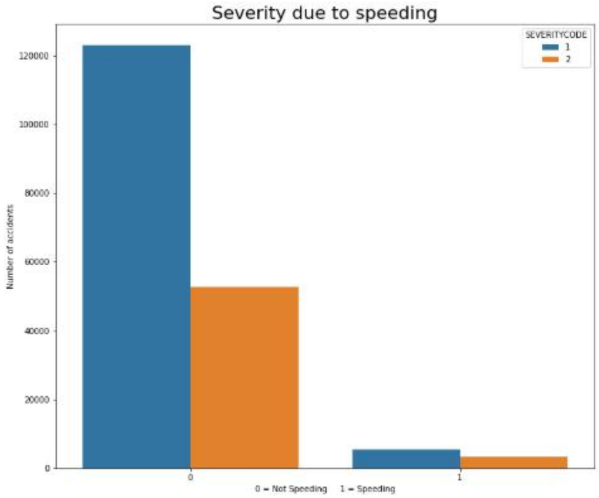
PERSONCOUNT, PEDCOUNT, PEDCLYCOUNT are directly related to the target variable, those are the 3 most correlated, followed by S*DOT_COLCODE, PEDROWNOTGRNT, ADDRTYPE* and *COLLISIONTYPE*.

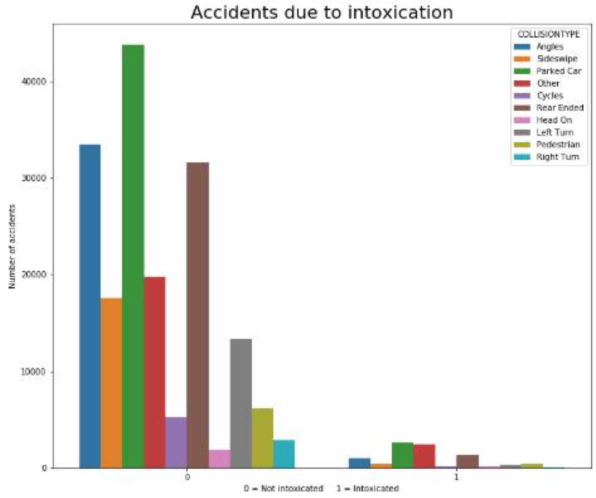


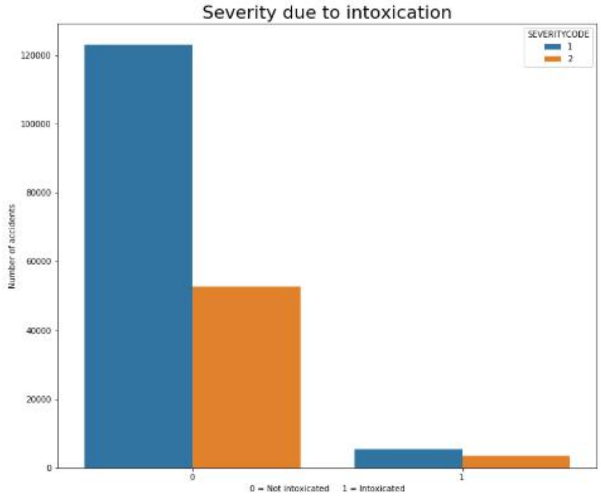


Inattention causes more accidents in Seattle than intoxication or speeding:



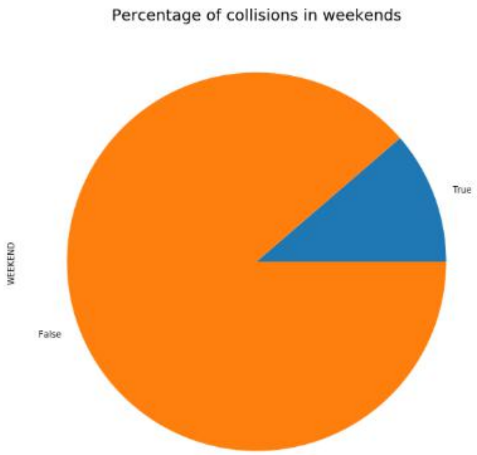


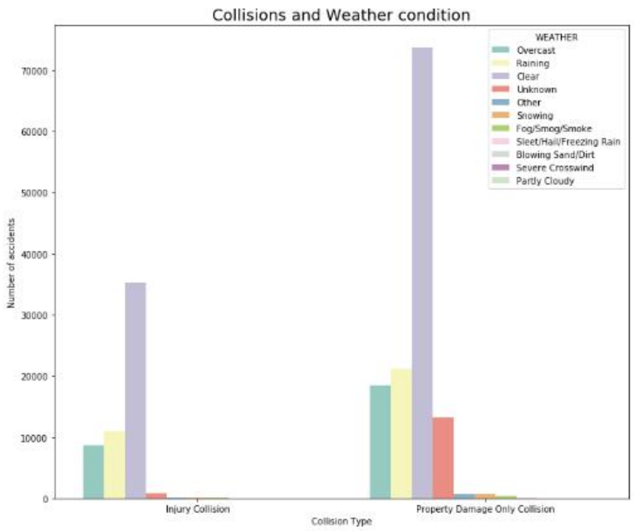




Observations

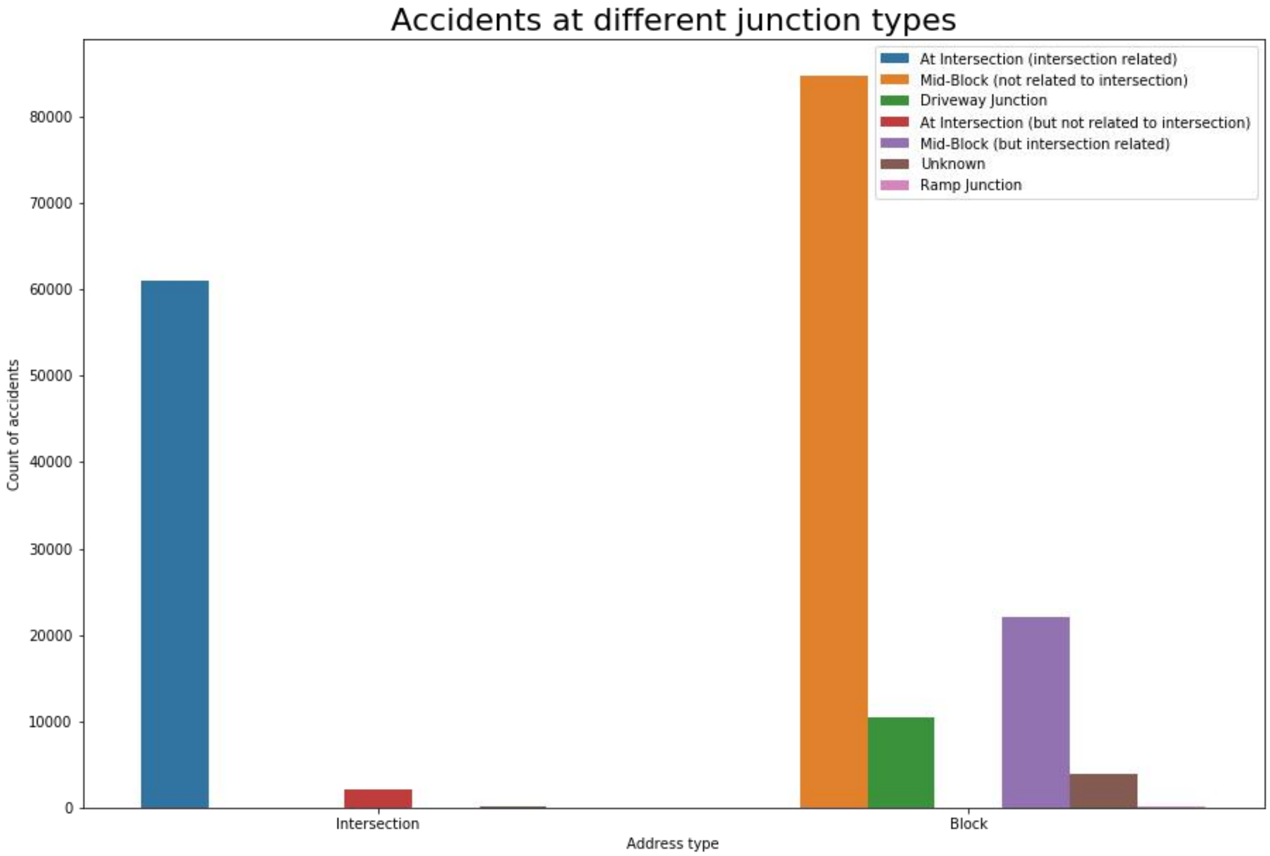
- Speeding and intoxication do not cause many accidents. Compared with other cities, it seems that Seattle has well controlled its system to avoid these types of problems.
- Most accidents are caused by negligence of drivers are simple inattention.
- Most collisions are against parked vehicles, the rear end and angles of the vehicles.





Observations

- Most accidents occur during weekdays.
- It seems that the weather is not a key factor. However, even though most accidents happen in clear weather, there are also many others during overcast and rainy days.



Observations

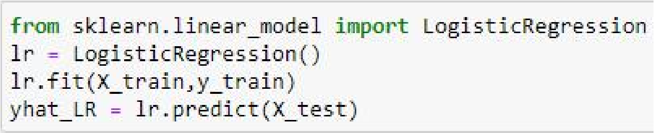
- After cleaning the data, the 751 Alley collisions were dropped.
- Most accidents occur at mid-block and intersections.

4. *Modeling*

After selecting the feature sets X and y and normalize X, it is time to use the test-train split and select a model to use. The distribution to the sets is: 80% train, 20% test.

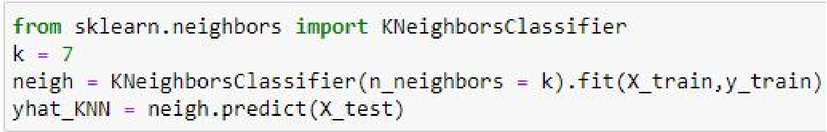
For our purpose, classification models are the best option, they can provide the probabilities of an accident may cause property damage or harm people. In order to find the best model, we selected 4 types: Logistic Regression, K-Nearest Neighbors, Decision Tree, Support Vector Machine (SVM) with the following structures.

4.1. Logistic Regression



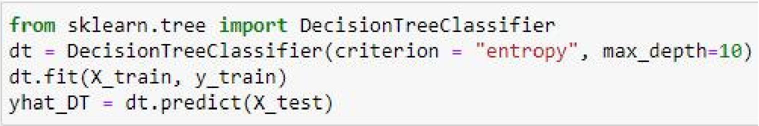
```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression()
lr.fit(X_train,y_train)
yhat_LR = lr.predict(X_test)
```

4.2. K-Nearest Neighbours



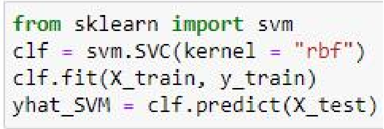
```
from sklearn.neighbors import KNeighborsClassifier
k = 7
neigh = KNeighborsClassifier(n_neighbors = k).fit(X_train,y_train)
yhat_KNN = neigh.predict(X_test)
```

4.3. Decision Tree



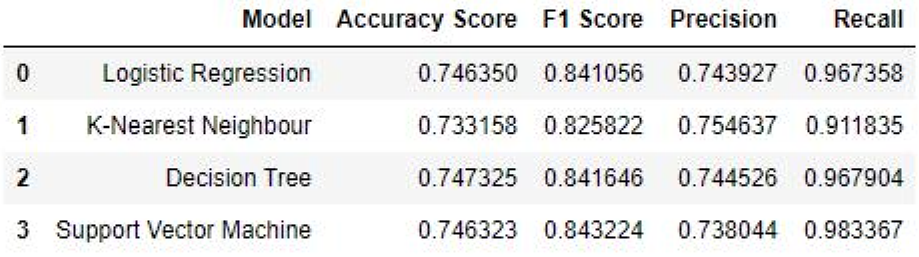
```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(criterion = "entropy", max_depth=10)
dt.fit(X_train, y_train)
yhat_DT = dt.predict(X_test)
```

4.4. Support Vector Machine



```
from sklearn import svm
clf = svm.SVC(kernel = "rbf")
clf.fit(X_train, y_train)
yhat_SVM = clf.predict(X_test)
```

4.5. Performance and Results



	Model	Accuracy Score	F1 Score	Precision	Recall
0	Logistic Regression	0.746350	0.841056	0.743927	0.967358
1	K-Nearest Neighbour	0.733158	0.825822	0.754637	0.911835
2	Decision Tree	0.747325	0.841646	0.744526	0.967904
3	Support Vector Machine	0.746323	0.843224	0.738044	0.983367

5. *Conclusions*

Discussion

- As we can see, all the selected models work similarly, from them we can tell the best was the Decision Tree, although it is likely that the KNN model may work better finding the optimal K.
- Also, as an extra, the Logistic Regression model performed well and, on its advantages, it could tell the probability for both classes of the target value.
- Obviously, these models can improve if the feature sets are handled better. For now, we can consider them good models for project purposes.

Final Observations

- Most collisions occur during day light on weekdays and commonly against parked cars.
- Most accidents happen at street intersections.
- Surprisingly, speeding and intoxication do not cause too many collisions in Seattle.
- On the other hand, inattention is the most common cause of an accident.
- Weather conditions do not play a significant role.

In this project, all the knowledge of the IBM Data Science Course has been tested.

I analyzed the relation between car accidents and different factors, such as: negligence of drivers, environmental issues, locations, people involved and so on. All of this in order to create a successful predictive model for 2 types of collisions (property damage or people injuries).

6. *Future Directions*

Undoubtedly, a model with these characteristics working with real time data and alert notifications (whether it is appropriate or not to drive under certain conditions) would significantly reduce crashes in many cities around the world. Additionally, this model serves to make all the population aware of the importance of being careful while driving, not only driving properly but also taking care of others on the road.