PREDICTING CAR ACCIDENT SEVERITY

1 - Introduction / Business Understanding

In general

I will enumerate here some data of interest about the problem that we are going to analyze/evaluate:

- Around 1.35 million people die each year as a result of traffic accidents.
- There is a clear problem in each country, and the objective is to achieve a maximum reduction in car or traffic accidents on public roads. Even the WHO sets ambitious goals such as reducing the number of deaths and injuries from traffic accidents worldwide by almost half.
- Traffic accidents cost most countries 3% of their GDP.
- More than half of deaths from traffic accidents affect "vulnerable road users", that is, pedestrians, cyclists, and motorcyclists.
- Despite the fact that low- and middle-income countries have approximately 60% of the world's vehicles, they account for more than 93% of deaths related to road accidents.
- Traffic accidents are the leading cause of death in children and young people between the ages of 5 and 29.

In particular for our case:

As for our business case, we are going to focus on the city of Seattle. Our audience in this case is the city council, its mayor and politicians want to solve the big problem of traffic accidents and seek analysis in the available data to be able to implement security measures and applications that help to reduce to the greatest extent possible traffic accidents and its severity.

2 - Data

Brief Description

We need to work on finding the severity of damage caused by accidents and for that we will require a large number of reports on traffic accidents so we can work on a prediction model. The data set provided for this exercise have around 195,000 accidents in the state of Seattle (location in which our focus will be for this exercise), from 2004 to the date it is issued (in 2020), and in which 37 attributes are recorded, such as location, collision type, date, weather, road conditions, etc. The dependent variable, SEVERITYCODE, contains numbers that correspond to different severity level caused by the accident. '1' indicates property damage only collision, and '2' indicates injury collision.

Data Source

These data have been collected and shared by the Seattle Police Department (Traffic Records) and we have access to the records using the given link.

Data set name

Data-Collisions

Usage

The data will be used so that we can determine which attributes are most common in traffic accidents in order to target prevention at which are the riskiest points or locations where they happened. We will be able to provide recommendations based on severity of the accidents so far (for what we have data), location of them, weather conditions in which these took place, etc

Data cleaning

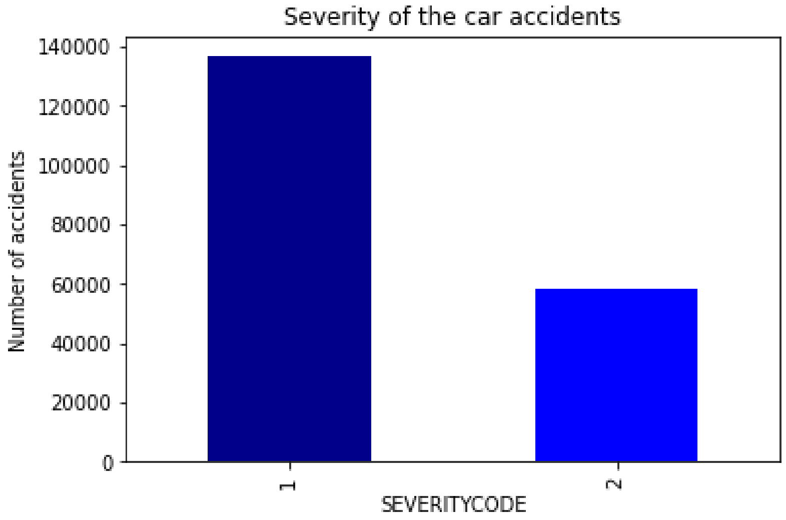
Before starting to work with the data, it has to be pre-process and cleaned so we use only data that it is interested or valid for the analysis we are going to perform. In this part, data will be cleaned to avoid having missing or unusual values. The goal is to have the data in a way that will be better and more effective for our study. We removed some columns which were not useful for our case study. Also categorical data was converted to numerical data in order to implement the machine learning algorithms.

3 - Metodology

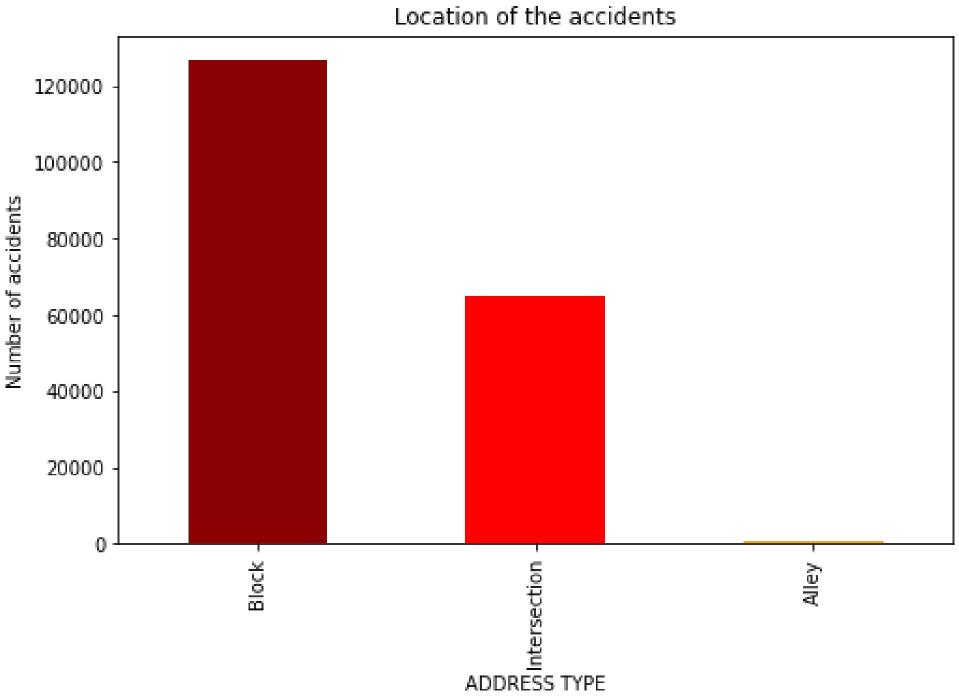
Studing the data

Some analysis has been done before choosing the variables to study and to see which of them will have more impact in the severity of an accident and in the number of accidents that occur.

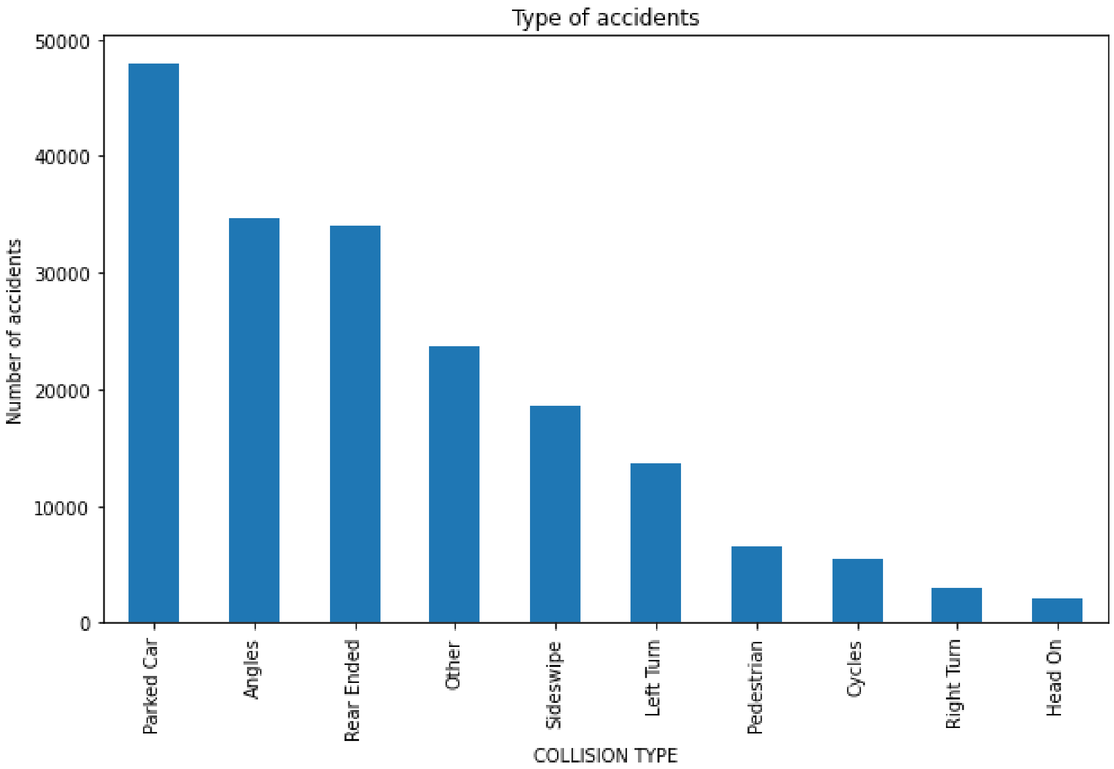
For example, it has being checked from the total amount of accidents that occur how many of them are severe or cause injures on people, what are the most common points for the accidents to happen, which kind of collision is more likely to happen, etc



Aprox. 1/3 of the accidents is causing some injure/damage on people. 2/3 are just property damage.



Aprox. 2/3 of the accidents are happening on blocks while 1/3 are on intersections. Alleys and rest have almost no incident..

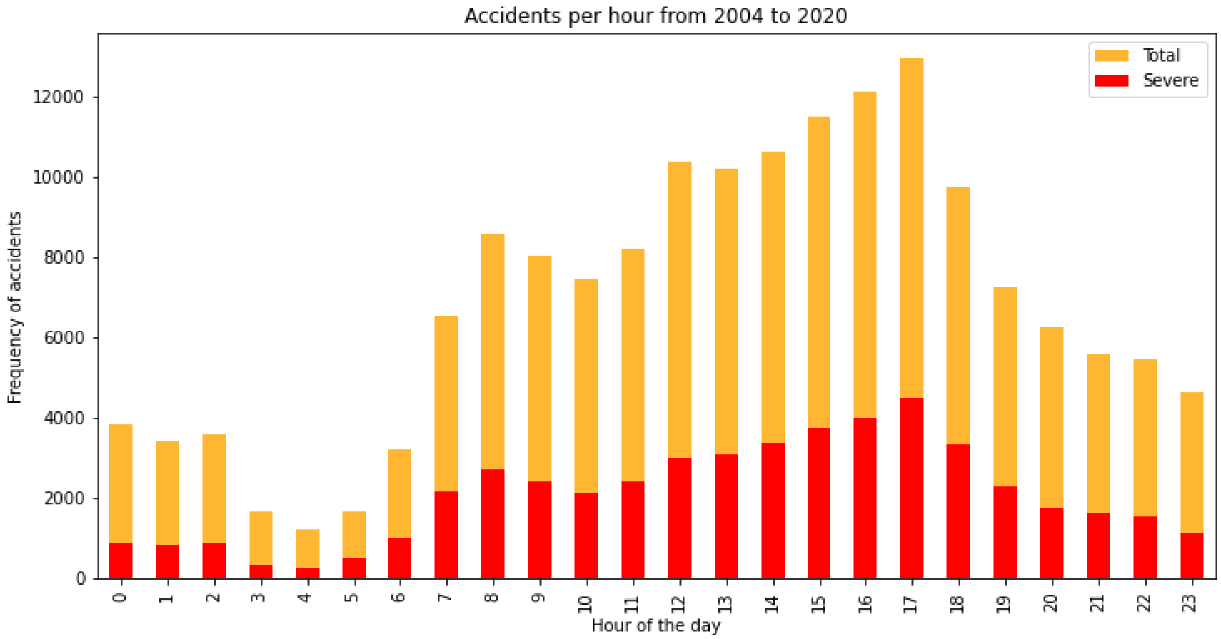


We can observe that most of the accidents are happening with parked cars. While parking (most probably) or passing by. In angles or in rear ended places of the car is it also quite often. We would need to suggest to the government to implement wider or safer parking places and to insist in reminding people to stay focus while driving as this seems to be due to distractions or lack of space.

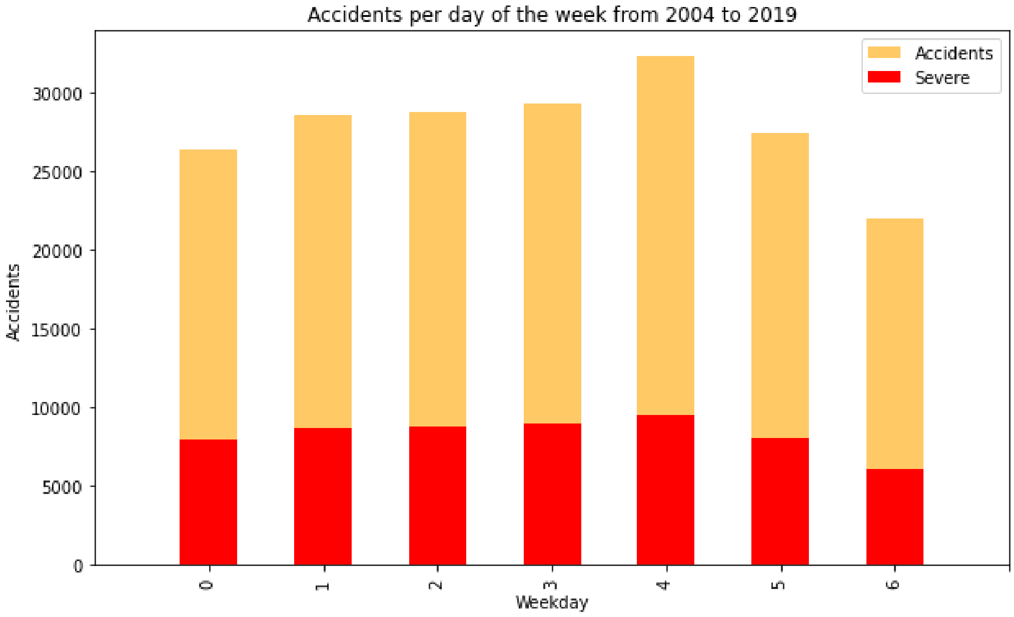
I have study other parameters and some of the conclusions are:

- Most accidents happen during daylight.
- Almost half of them are on mid-blocks and the other half have something to do with intersections.
- Only around 15% of the times it is related to inattention
- More than half of the times, aprox. 60% weather has not an impact on the accidents. It does not seem to be a high-risk factor or the main issue.
- Most of the accidents are in dry roads. This condition does not seem to be a high-risk factor or the main issue. Similar to the weather conditions.
- Most of the accidents are during the day.
- Only a few times (aprox 4%) it happens due to high speed.

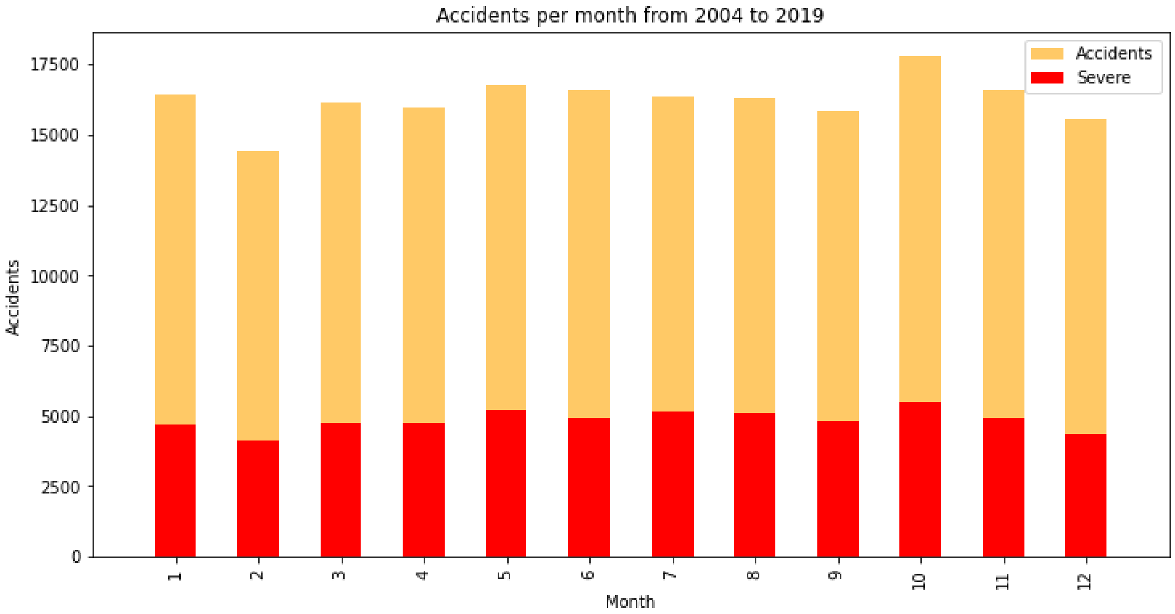
Let’s also check the Trend and the Seasonality of the accidents:

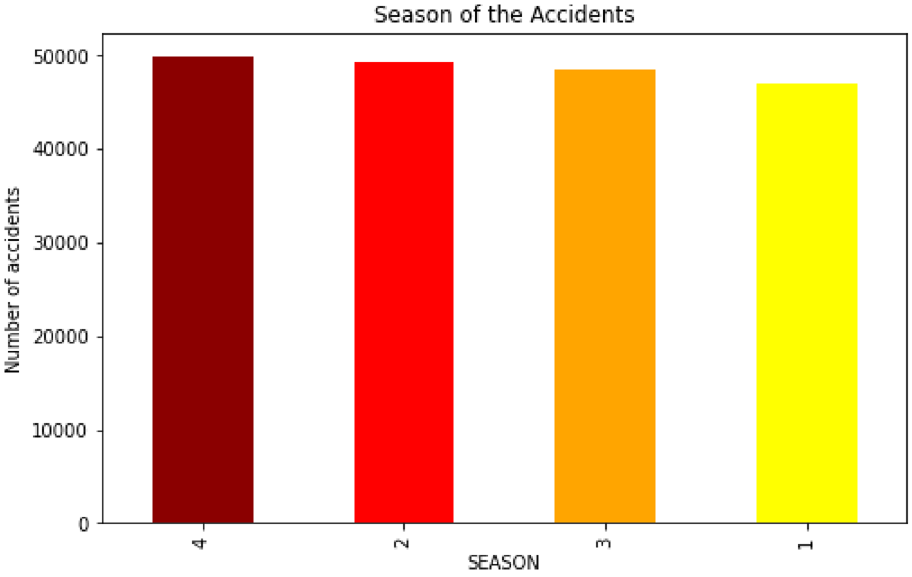


We can see that the pick is around working hours (coming back home from work). During night hours and early morning, it is not an issue. Measures will have to be implemented on this main range of hours.

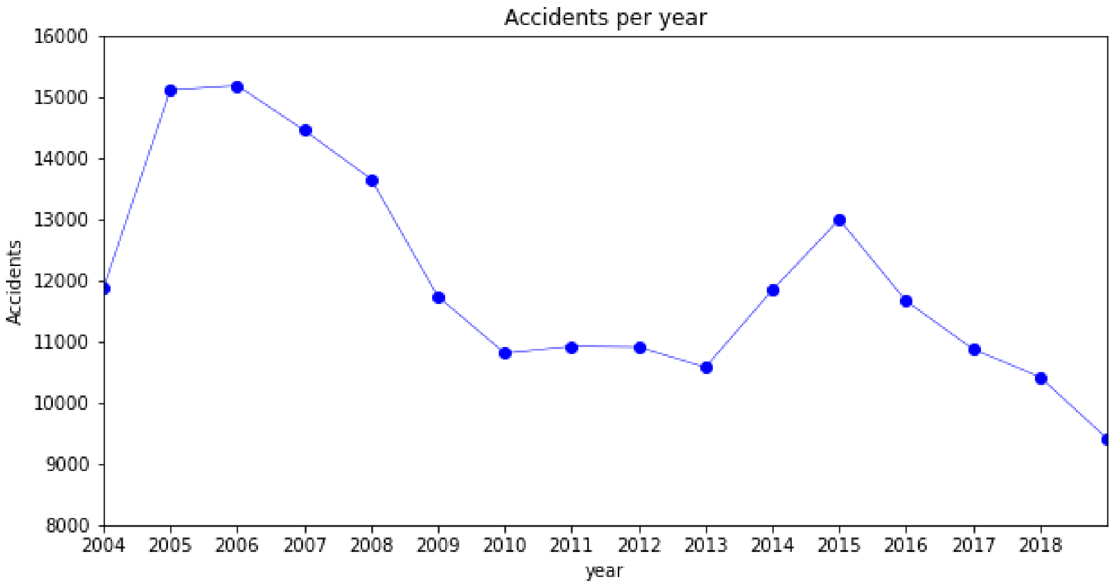


It is quite balanced comparing the days of the week and we can see it is better specially on Sunday (not working day). Friday is the highest probably due to the nightlife and traveling for the weekend.





Seasonality does not show any big different between months or seasons.



We can see that the trend of the last years is that accidents are decreasing. Measurements from politicians as well as TV and awareness of people is making this possible.

Data Preparation

All the rows that are not necessary in order to build the predictive model are dropped. The data is going to be converted from categorical to numerical values for it to be studied. The final data frame for study will be built and we will compare the accuracy and results of the machine learning techniques used.

4 - Results

Algorithm	Jaccard	f1-score	Precision	Recall
Logistic Regression	0.72	0.66	0.69	0.62
KNN	0.73	0.68	0.71	0.73
SVM	0.72	0.66	0.69	0.72

For our case study precision is the % of predicted severe (ID=2) accidents that will cause some injure or damage on people. The Recall instead, is the % of the severe accidents that were properly predicted. Therefore, is more important to take into consideration the Recall as it will help the authorities to be prepared for when a severe accident might occur

The Logistic Regression, KNN, and SVM models have similar accuracy, however the time that took me to run SVM was much higher. Seeing the Recall value and the time it took to run, I suggest the **KNN method to be used**.

5 - Discussion

Several methods can be used to predict severity of the accidents, also, to check when the possibility of an accident to happen will be higher. I have used several graphics to check which kind of accidents, places of them, and different situations can affect the number of the accidents that will occur and its severity. I have also checked and compared building different predictive models which one have more accuracy and it is more suitable for the variables that by decision, I think they can affect more the accident to happen and the severity of it. From all the solutions and cases studied I believe KNN will be the best method to be used. It will help authorities to be prepared when an accident which will cause injures on people will happen.

6 - Conclusion

In a World where the need of using transport is a mut for almost all of us, accidents might occur often. Specially during certain conditions. It is important to help authorities to know when an accident is more likely to occur to at least try to implement measures during the worst days or days when accidents to happen has a higher probability. Also, creating some model /algorithm that can predict the severity of the accidents will help them to be prepared for such cases. Knowing the locations, times and conditions in which this accident happened and the probability of being severe will help them to know before even going to the place of the events to be ready to provide attention faster as well as to promote market campaigns to make people aware of this problem and the days, places, conditions that are worst, so they might decide change the plan or be more focus during those moments.