**IBM CAPSTONE**

**1. Introduction**

**1.1. Background**

Traffic accidents are a significant source of deaths, injuries, property damage, and a major concern for public health and traffic safety. They involve a great deal of human exposure and economic interest. Only in the US, 4.4 million people are injured every year in car accidents, and over 38.000 people die on the road. This implies a total cost for US citizens of $871bn, of which $380bn belongs to medical expenses.[[1]](#footnote-1) Avoiding car accidents might be almost impossible, but understanding and reducing severity risk can have benefits in terms of health and less economic impact of governments, health institutions and insurance companies.

**1.2. Problem**

Data can contribute to understanding the drivers of severity of car accidents and help identify patterns in order to develop preventive measures. Property damage vs physical injuries have different impacts in diverse fields, such as legal treatment, driver health and economic compensation.

Bear in mind however, fatalities are out of the scope of this study.

**1.3. Target audience**

Several stakeholders can be interested in such an analysis. Firstly, administrations would be able do city planning and distribute resources more efficiently depending on the areas of highest accident severity probability.

Secondly, data insights would help car manufacturers develop new technologies and components to minimize risk factors and improve the safety of the vehicle.

Moreover, insurance companies would benefit from a model that differentiates between personal and property damage. Having a better understanding of each type will help on risk assessment, policy amount and client targeting.

Finally, drivers themselves would be able to take precautions by understanding possible risk factors and plan their travels in a way to maximize their safety.

**2. Data Description and Cleaning**

The data I used is the CSV file provided by IBM Capstone in Coursera. This database, called “Collisions” and published by SDOT Traffic Management Division, contains a total 37 attributes related to severity of accidents (our prediction target), incident location, road state and number of agents involved, among others. Data ranges from 2004 to 2019 and includes 194,673 samples.

**2.1. Feature Selection**

However, many of these attributes were redundant or couldn’t be included in the model as they were consequences of the accident, rather than a possible driver:

|  |  |  |
| --- | --- | --- |
| Reason for dropping | Redundant index referring to another Feature | Consequence of the incident, cannot be input to predict severity |
| Dropped Features | COLDETKEY  INTKEY  LOCATION  SDOT\_COLCODE  SDOTCOLNUM  ST\_COLCODE  SEGLANEKEY  CROSSWALKKEY  EXCEPTRSNCODE  EXCEPTRSNDESC | INJURIES  FATALITIES |
| Kept Features | OBJECTID  ADDRTYPE  SEVERITY  SEVERITYDESC  COLLISIONTYPE  PERSONCOUNT  PEDCOUNT  PEDCYLCOUNT  VEHCOUNT  INCDATE | UNDERINFL  WEATHER  ROADCOND  LIGHTCOND  SPEEDING  HITPARKEDCAR  INCDTTM  JUNCTIONTYPE  SDOT\_COLDESC  INATTENTIONIND |

For more info on the metadata, please see Annex 1.

Note: in this case severity was restricted to “property damage” and “physical injury”.

**2.2. Data Cleaning (Methodology I)**

- NaN: Deleted all NaN rows, since dataset was large, and it was considered preferable to keep quality data.

- Binarization: Transformed Yes/No cells into binary numbers, as well as severity.

- Pearson correlation: after previous Feature selection, no other features had a large enough index to prove relevant redundancies.

**2.3. How will the data solve the problem?**

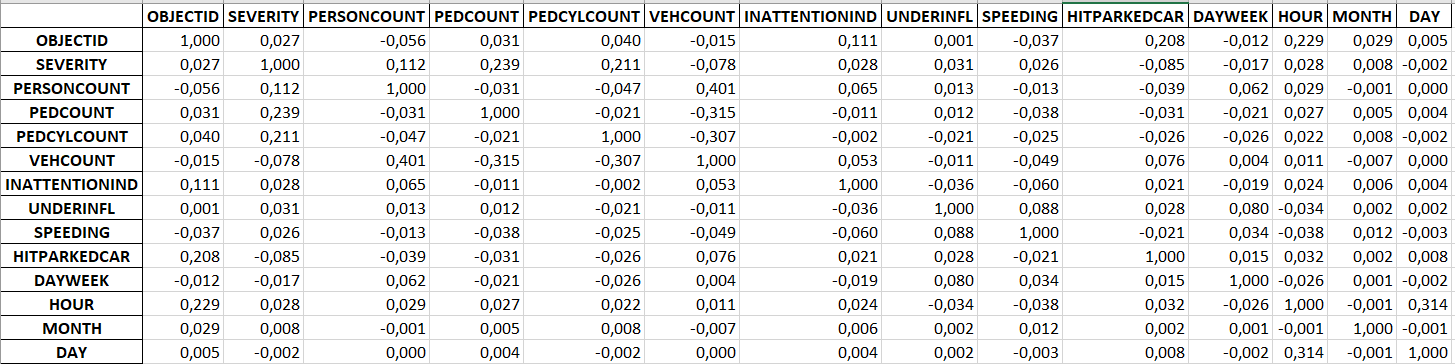
With this revised dataset, I will study which are the most suitable features to predict the target variable: SEVERITY. Independent variables include information about the road’s state, location of the incident, amount of vehicles/people involved and other factors as well as driver factors such as speeding or presence of alcohol.

Through the incoming study, I will create a model that will explain which of the previous 20 variables are the most determinant factors regarding accident severity. By determining these factors, it will be possible for stakeholders to develop preventive measures, ensuring fast action from health institutions and allocating resources in a more efficient manner.

**3. Exploratory Data Analysis (Methodology I)**

- Numerical features

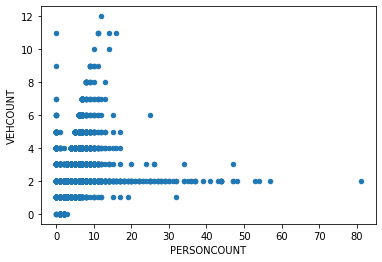
After reducing the number of features, I analyzed possible correlations between variables with the aim to reduce redundancies. However, the remaining variables had almost no correlation between one another, with the highest correlation p < 0.3.



Data imported to Excel from Python Notebook.

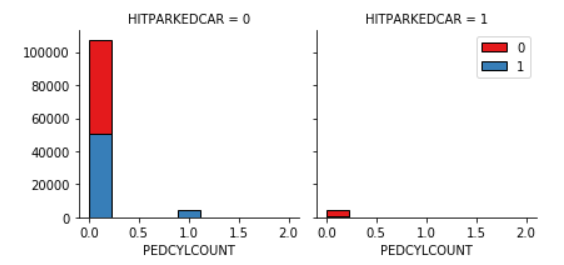
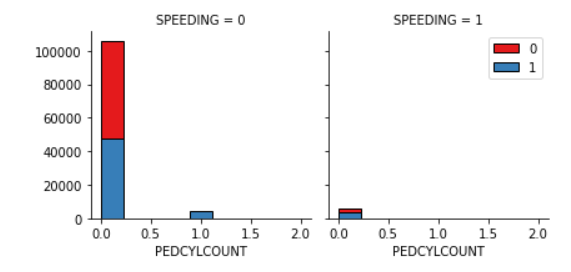
The target variable “SEVERITY” only correlated to a minor extent with “PEDCOUNT” and “PEDCYLCOUNT” (amount of pedestrians and cyclists, respectively). Since the amount of bicycles doesn’t seem like a good predictor (before an accident it’s difficult to know the amount of cyclists in the area), I decided not to include it in the future models.

Moving on to “VEHCOUNT” (amount of vehicles) I noticed that there’s a certain correlation between this variable and PERSONCOUNT (amount of people in vehicles): the more vehicles involved, it’s more likely that more people will be involved in an accident. This one could also be a good predictor candidate.

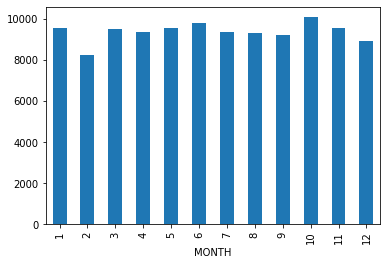
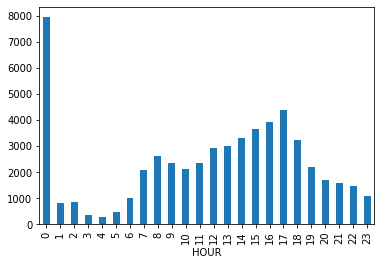


I also checked If there were redundancies between PEDCOUNT and PERSONCOUNT. They only had a p = -0.031, which implies no linear relationship. I also checked the graph and couldn’t find any obvious pattern. Therefore, it was OK to include both variables in the model.

I also thought I would be a good idea to include other factors which were not related to the amount and kind of people. Therefore, I considered variables such as SPEEDING and HITPARKED CAR. When speeding is involved, cases tend to be more severe, whereas if a parked car is hit, in many cases it’s only considered property damage.

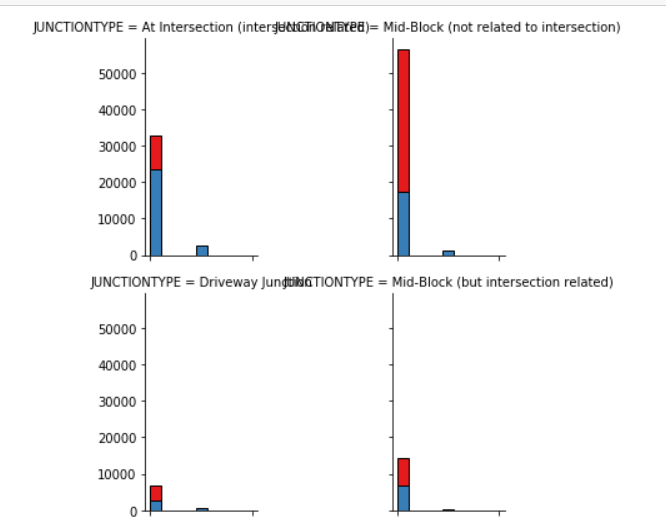
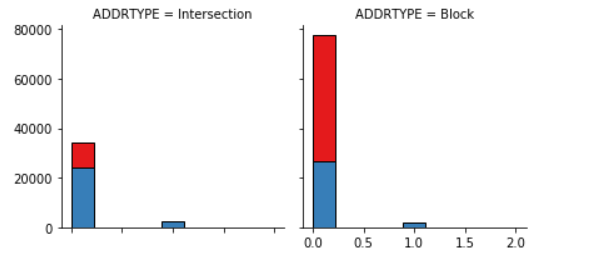


I also created two new variables based on the time of the incident: HOUR and MONTH. Although the month of the incident didn’t prove to show any pattern, one could clearly see that most accidents took place during the day, at rush hour in the afternoon (let’s ignore 0:00h since some registers include 0:00 if the hour is unknown).

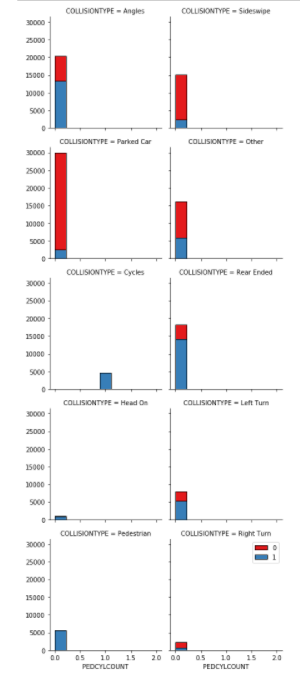


* Categorical features

I also wanted to analyze some categorical features, as they included different information that could be useful. For instance, the location of the incident provided us some information: more severe accidents (blue) are more likely to happen in an intersection than just property damage.



It is also interesting to notice how the COLLISIONTYPE can also help determine the severity of the accident (blue = injury, red = property damage).

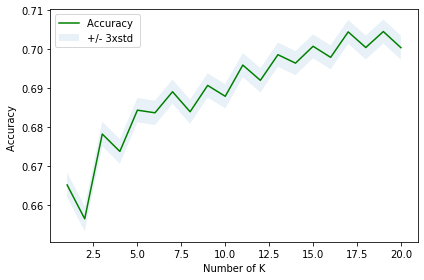


**4. Model creation (Methodology III)**

With this information now I could create and train various models to try to achieve an optimal accuracy. I selected the variables "PERSONCOUNT", "PEDCOUNT", "UNDERINFL", "HITPARKEDCAR", "HOUR", "ADDRTYPE", "JUNCTIONTYPE" and "COLLISIONTYPE". The last 3 variables, since they were categorial, I had to carry out one hot encoding method to be able to train the model.

Afterwards, I cleaned the data using sample reduction to achieve a balanced dataset and got 110,946 high quality samples, without NaN.

Finally, I divided the data into train and test sets (20% test size) and trained 4 algorithms: KNN, Decision Tree, Logistic Regression and SVM. For KNN, I plotted the different Ks to see which one had the best accuracy (in this case 17).



**5. Results**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Accuracy Score | Jaccard | F1 Score | Log Loss |
| KNN | 0.70 | 0.70 | 0.70 | NA |
| Decision Tree | 0.70 | 0.66 | 0.64 | NA |
| SVM | 0.71 | 0.71 | 0.71 | NA |
| LR | 0.70 | 0.70 | 0.70 | 0.55 |

As we can observe in the previous table with some evaluation scores for each algorithm, they perform very similarly. SVM is the most accurate model by 0.01 (even though it takes quite longer to process), while the Decision Tree algorithm doesn’t have such a high accuracy as its counterparts.

**6. Discussion**

Based on information related to amount of people in vehicles, location of an accident, accident type and other driver conditions (e.g. speeding, alcohol intake) we can eliminate part of the variance of the dataset. However, the accuracy is of only a 70% despite including a relevant amount of features of different types.

We can draw the conclusion that more information is required to predict the severity of an accident. Some future paths for further investigation could be:

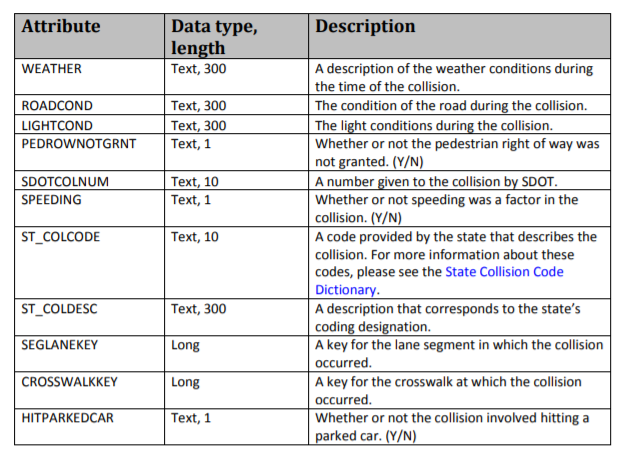
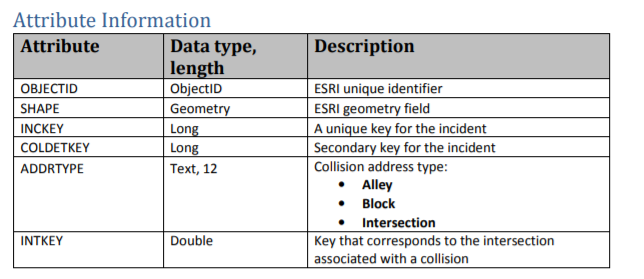
* Vehicle conditions: brand, type, age, etc.
* Driver conditions: age, physical state, previous driving history, time since obtained license, etc.
* Insurance type: perhaps vehicles with more extense policies are more prone to suffer more severe accidents due to less attention of the driver.

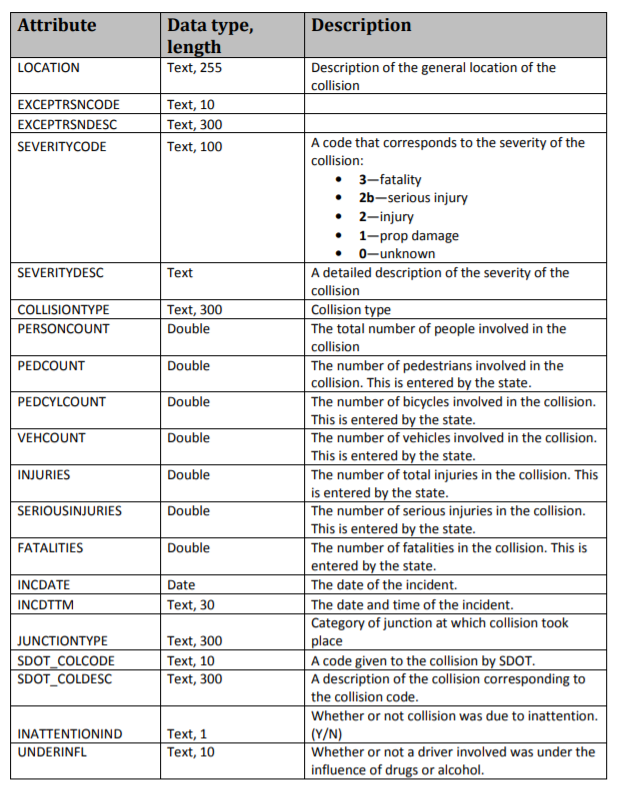
**7. Conclusions**

In this study, it was studied how the severity of an accident could be determined based on different factors present in traffic incidents. The objective was to understand the weight and interactions of these factors to help diverse stakeholders (public administrations, health institutions and insurance companies) reduce human and economic costs associated with car accidents. I identified as most relevant factors: the amount of people in the vehicle, location of the accident and type of intersection, accident type, hour of the accident and as circumstantial factors of the driver the presence of speeding and alcohol intake.

The models I developed, with an accuracy of around 70%, with SVM as the best performer, will help stakeholders understand some of the determinant factors of accident severity. However, there’s much more information we are missing: vehicle conditions, driver conditions or insurance type for example. This data could be very useful and might be worth analyzing in future studies.

**ANNEX 1**





1. Association for Safe International Road Travel <https://www.asirt.org/safe-travel/road-safety-facts/#:~:text=Annual%20United%20States%20Road%20Crash,enough%20to%20require%20medical%20attention.> [↑](#footnote-ref-1)