Coursera - Applied Data Science Capstone
Ana Cristina Bittar de Oliveira

Report – Week 2

1. Introduction

How to prevent or, at least, reduce the severity of road accidents is a very important question. A better understanding of the probable causes of collisions could be a good approach in order to suggest actions to be taken or improvements to be made in traffic engineering systems willing to prevent them.

The main potential consequences of a car collision are:

- Injuries
- Traffic jams
- Related costs (government, insurance companies, individuals, *et al*)

Seattle Traffic Management Division releases weekly updates about all collisions recorded by Traffic Records since 2004. The main goal is to analyze that data to identify patterns and make predictions about risk and severity of the accidents based on some common attributes. With this information an alert system could be developed to provide guidance on what to do in short, medium, or large term.

Some examples of actions to be taken could be:

- deviate traffic in pre-defined times or weather conditions (preventive)
- improve traffic signs
- reduce speed in specific roads in determined weather conditions
- reduce speed in specific roads (permanently)
- make structural changes on traffic
- educative campaigns and advertisement
- recommendations to improve the data collection
- other

With limited resources the city must decide which actions will be taken, and the expected severity of the collisions might be the key to allocate resources.

The present work involves analysis of data related to collisions in the city of Seattle. It would be out of scope to generalize this model for sites other than those included on the data to be analyzed, although intuition indicates that, for similar traffic systems, the results could be extended. The expected result is a model capable to predict an accident severity in pre-determined conditions.

The data will be used to make predictions about probability and severity of car accidents based mainly on:

- Weather conditions
- Road conditions
- Light conditions

Further analysis (maybe using additional datasets) could be made to check if it's possible to correlate some driver conditions (as speed, influence of drugs or alcohol, inattention) with other attributes like time of the day, day of the week, events like concerts or games (if the data is available) and define whether or not those attributes could be used for predictions. This is out of scope of the present report.

Other aspects that should be observed for future improvements are:

- There's very detailed geographic information, but no demographic about the conductors. That could be valuable information in case of educative campaigns, for example.
- The data went back 15 years, and traffic volume increases rapidly along the years. But it was not taken into consideration for the model

First attempt was an unsupervised model but its results were not good. The final model used a Decision Tree Classifier machine learning alghoritm.

2. Data

The modelling was based on the datasets related to collisions in the city of Seattle, provided by SPD and recorded by Traffic Records.

The CSV file was obtained directly from Seattle Open Data Portal.
https://opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv

The dataset provides information about all types of collisions occurred in the city of Seattle since 2004 and is updated weekly.

The city also provides metadata for better understanding of the dataset:

Attribute Information:
https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf

Some attributes are redundant, or the number of observations is irrelevant, and were excluded (the numbers and statistics refer to the CSV file extracted on Sep 13th).

```
Info about the original data file:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 221389 entries, 0 to 221388
Data columns (total 40 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   X                213918 non-null  float64
 1   Y                213918 non-null  float64
 2   OBJECTID         221389 non-null  int64
 3   INCKEY           221389 non-null  int64
 4   COLDETKEY        221389 non-null  int64
 5   REPORTNO         221389 non-null  object
 6   STATUS           221389 non-null  object
 7   ADDRTYPE         217677 non-null  object
 8   INTKEY           71884 non-null   float64
 9   LOCATION         216801 non-null  object
 10  EXCEPTRSNCODE    100986 non-null  object
 11  EXCEPTRSNDESC    11779 non-null   object
 12  SEVERITYCODE     221388 non-null  object
 13  SEVERITYDESC     221389 non-null  object
 14  COLLISIONTYPE    195159 non-null  object
 15  PERSONCOUNT      221389 non-null  int64
 16  PEDCOUNT         221389 non-null  int64
 17  PEDCYLCOUNT      221389 non-null  int64
 18  VEHCOUNT         221389 non-null  int64
 19  INJURIES         221389 non-null  int64
```

```
 20  SERIOUSINJURIES  221389 non-null  int64
 21  FATALITIES       221389 non-null  int64
 22  INCDATE          221389 non-null  object
 23  INCDTTM          221389 non-null  datetime64[ns]
 24  JUNCTIONTYPE     209417 non-null  object
 25  SDOT_COLCODE     221388 non-null  float64
 26  SDOT_COLDESC     221388 non-null  object
 27  INATTENTIONIND   30188 non-null   object
 28  UNDERINFL        195179 non-null  object
 29  WEATHER          194969 non-null  object
 30  ROADCOND         195050 non-null  object
 31  LIGHTCOND        194880 non-null  object
 32  PEDROWNOTGRNT    5192 non-null    object
 33  SDOTCOLNUM       127205 non-null  float64
 34  SPEEDING         9928 non-null    object
 35  ST_COLCODE       211976 non-null  object
 36  ST_COLDESC       195159 non-null  object
 37  SEGLANEKEY       221389 non-null  int64
 38  CROSSWALKKEY     221389 non-null  int64
 39  HITPARKEDCAR     221389 non-null  object
dtypes: datetime64[ns](1), float64(5), int64(12), object(22)
memory usage: 67.6+ MB
```

2.1. Data Preparation

Unnecessary or irrelevant attributes (few observations) were removed. A complete list is below:

```
['OBJECTID','INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'INTKEY', 'LOCATION',
'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'INCDATE', 'SDOT_COLCODE',
'SDOT_COLDESC','SDOTCOLNUM', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY',
'CROSSWALKKEY','INATTENTIONIND',
'PEDROWNOTGRNT', 'SPEEDING']
```

The below shows statistics about some numerical variables related to severity. In a supervised model those attributes can be used to define different weights for severity scale. For example, a collision with 41 serious injuries is more significant than one with 2 serious injuries, although they get the same severity code.

	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INJURIES	SERIOUSINJURIES	FATALITIES
Count	221,389	221,389	221,389	221,389	221,389	221,389	221,389
Mean	2.227161	0.038136	0.027350	1.731057	0.373962	0.015209	0.001685
Std	1.470190	0.201815	0.164508	0.829259	0.732158	0.158072	0.044701
Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	2.0	0.0	0.0	2.0	0.0	0.0	0.0
50%	2.0	0.0	0.0	2.0	0.0	0.0	0.0
75%	3.0	0.0	0.0	2.0	1.0	0.0	0.0
Max	93	6	2	15	78	41	5

Another transformation was to replace categorical for Numeric or Boolean values.

Rows with missing values for the main attributes (WEATHER, ROADCOND, LIGHTCOND) were removed, as were observations with SEVERITYCODE = ‘0’ or SEVERITYCODEDESC = ‘Unknown’.

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND
Count	194699	194699	194699	194699
Unique	5	11	9	9
Top	1	Clear	Dry	Daylight
Freq.	133575	114556	128304	119384

Final Dataframe Information:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 194697 entries, 0 to 221388
Data columns (total 19 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   X                 189455 non-null  float64
 1   Y                 189455 non-null  float64
 2   ADDRTYPE          194697 non-null  int64
 3   SEVERITYCODE      194697 non-null  object
 4   COLLISIONTYPE     194697 non-null  int64
 5   PERSONCOUNT       194697 non-null  int64
 6   PEDCOUNT          194697 non-null  int64
 7   PEDCYLCOUNT       194697 non-null  int64
 8   VEHCOUNT          194697 non-null  int64
 9   INJURIES          194697 non-null  int64
 10  SERIOUSINJURIES   194697 non-null  int64
 11  FATALITIES        194697 non-null  int64
 12  INCDTTM           194697 non-null  object
 13  JUNCTIONTYPE      194697 non-null  int64
 14  UNDERINFL         194697 non-null  bool
 15  WEATHER           194697 non-null  int64
 16  ROADCOND          194697 non-null  int64
 17  LIGHTCOND         194697 non-null  int64
 18  HITPARKEDCAR      194697 non-null  bool
dtypes: bool(2), float64(2), int64(13), object(2)
memory usage: 27.1+ MB
```

2.2.Metadata

Data Set Summary

Data Set Basics	
Title	Collisions—All Years
Abstract	All collisions provided by SPD and recorded by Traffic Records.
Description	This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.
Supplemental Information	
Update Frequency	Weekly
Keyword(s)	SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle
Contact Information	
Contact Organization	SDOT Traffic Management Division, Traffic Records Group
Contact Person	SDOT GIS Analyst
Contact Email	DOT_IT_GIS@seattle.gov

Attribute	Data type, length	Description
LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • **3**—fatality • **2b**—serious injury • **2**—injury • **1**—prop damage • **0**—unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text, 300	Collision type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.
INJURIES	Double	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	Double	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	Double	The number of fatalities in the collision. This is entered by the state.
INCDATE	Date	The date of the incident.
INCDTTM	Text, 30	The date and time of the incident.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.
SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.

Attribute	Data type, length	Description
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.
SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)

3. Methodology

Apply machine learning alghoritms to the available data to find the best alghoritm to make predictions about severity of car accidents.
Depending on the results iteratively work on data trying to find arrangements to best fit the model.

4. Results

In this work Decision Tree Classifier machine learning algorithm was applied to predict severity levels of car collisions based on provided attributes.

Severity levels provided were (records with severity code 0 = Unknown were not considered):

3—fatality
2b—serious injury
2—injury
1—prop damage

The first attempt is to test the algorithm capability of predicting severity. Its basic form with default parameters was used:

```
DecisionTreeClassifier(criterion='entropy', max_depth=4)
```

and initially the three main features (WEATHER, LIGHTCOND and ROADCOND):

Using 60% of the sample for training and 40% for testing, the result was an initial accuracy of 69%.

With that result the efficacy of Decision Tree Classifier machine learning algorithm was proved for the case study, and the next steps were improvements to achieve better accuracies.

4.1. Including features

Using all the attributes applicable, with a training set of 70% of the samples the accuracy increased to 73%.

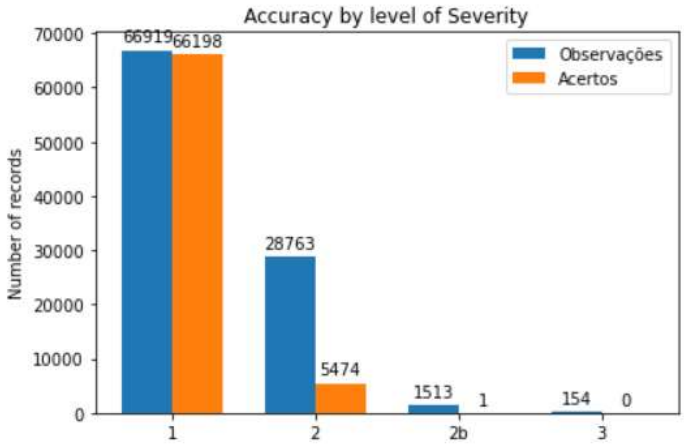
The attributes used were:
`['ADDRTYPE', 'COLLISIONTYPE','JUNCTIONTYPE','WEATHER', 'ROADCOND', 'LIGHTCOND', 'HITPARKEDCAR']`

Those are the only applicable attributes once the counting of injuries, fatalities, e.a. can't be used for predictions.

4.2. Changing parameters for the algorithm

Increasing the number of nodes the overall accuracy increased:
max_depth = 6 => accuracy 74%

At this point, with a good level of accuracy the next attempt was to check the accuracy for each level of severity, and the results were disappointing:



That was probably due to the small number of observations for the higher severity levels (imbalanced dataset).

4.3. Balancing the Dataset

As an attempt to improve the accuracy some changes were made manually. Part of records with severity code '1'and '2' were discarded (undersampling). Based on the number of injuries, severe injuries, fatalities and person involved the records were replicated to find some balance (oversampling). Unfortunately the level of accuracy for high severity incidents didn't raise.

5. Discussion

The dataset provides a lot of information, and it goes 15 years back, which is good. On the other hand it presents redundancies and features with an insignificant number of observations, probably because the data collected changes with time.

That is probably why it took a lot of time to analyze and clean the data before any alghoritm could be applied.

The main issue found was due to imbalanced data.

6. Conclusion

The choice of the machine learning algorithm for the given problem and available data is the main point. In the present case, the Decision Tree Classifier was used considering its similarity with human thinking in classifying based on simple decisions, and good computational performance.

There's a lot of external factors that can influence the severity of a car crash. The term itself is very subjective. For this case an accuracy of 74% on predictions is a good result when considering the main goal of estimating severity in order to better allocate resources to preventive actions.

However, since data was imbalanced it is necessary to correct this in order to get better accuracy for all severity levels. Despite many attempts of balancing the data the accuracy for high severity levels couldn't be improved.