Coursera - IBM Data Science Capstone

Predicting Accident Severity

Car accident severity (Week 1)

Rousseau, Antoine
10-20-2020

Contents

Problem

Traffic accidents are an endemic problem in the US. In 2018 alone, 12 million vehicles were involved in a car crash[1]. Individuals take the road, often out of necessity, not thinking twice about the conditions they will face, and how much they will put themselves and others at risk. **What if we could predict the severity of an accident, if one were to occur, based on time, date and environmental data, and thereby provide information that could reduce their frequency?**

As a result, authorities could optimize the allocation of rapid-response services and signal the risk of taking the road to the population in real-time through billboards. Car manufacturers could signal this information in the navigation system of the cars. Individuals could have access to an app that would suggest a higher risk of a severe accident if they were to take the road.

Our analysis is built upon the Seattle Collisions dataset. It contains, in tabular form, 31 attributes detailing each of the 194 673 collisions that took place in the Seattle from 2004 to today. Each entry is labelled according to severity. We will start by describing the dataset, gauging the quality of the data, and assessing the relevance to the problem of the available attributes. We will then transform it, selecting the useful attributes, addressing missing or unknown entries, and encoding it. We will then run several classification algorithms and evaluate their performance. Finally, we will discuss the findings and assess how the model answers the problem statement.

Data Understanding

The first step in analysing the data is to get an understanding of the attributes. Reading the Metadata pdf, we can allocate the attributes in the following buckets:

- **ID**: OBJECTID, INCKEY, COLDETKEY, REPORTNO, STATUS
- **Location**: X, Y, ADDRTYPE, INTKEY, LOCATION, JUNCTIONTYPE, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR
- **Unknown**: EXCEPTRSNCODE, EXCEPTRSNDESC
- **Accident description**: SEVERITYDESC, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC
- **Time/Date**: INCDATE, INCDTTM
- **Driver Behavior/State**: INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING
- **Environmental**: WEATHER, ROADCOND, LIGHTCOND

To answer our problem, we are interested in the predictive ability of the Time/Date and Environmental attributes and will therefore discard all other attributes in the data preparation stage.

We print the table to see how the data frame looks like:

[1] https://www.statista.com/topics/3708/road-accidents-in-the-us/

	SEVERITYCODE	INCDATE	INCDTTM	WEATHER	ROADCOND	LIGHTCOND
0	2	2013/03/27 00:00:00+00	3/27/2013 2:54:00 PM	Overcast	Wet	Daylight
1	1	2006/12/20 00:00:00+00	12/20/2006 6:55:00 PM	Raining	Wet	Dark - Street Lights On
2	1	2004/11/18 00:00:00+00	11/18/2004 10:20:00 AM	Overcast	Dry	Daylight
3	1	2013/03/29 00:00:00+00	3/29/2013 9:26:00 AM	Clear	Dry	Daylight
4	2	2004/01/28 00:00:00+00	1/28/2004 8:04:00 AM	Raining	Wet	Daylight

We describe the data:

```
RangeIndex: 194673 entries, 0 to 194672
Data columns (total 6 columns):
SEVERITYCODE    194673 non-null int64
INCDATE         194673 non-null object
INCDTTM         194673 non-null object
WEATHER         189592 non-null object
ROADCOND        189661 non-null object
LIGHTCOND       189503 non-null object
```

As we can see, the two Time/Date attributes are not of the “datetime” type nor the Environmental attributes of the “category” type. This will have to be addressed in the data preparation stage. We can also see that several rows from the Environmental attributes are missing (there are less than 194673 values for WEATHER, ROADCOND and LIGHTCOND)

The next step is checking the quality of the data starting with missing values. Checking the number of rows with at least one null reveals that less than 3% of them contain missing values. Additionally, after inspection of the unique categorical values of WEATHER, ROADCOND and LIGHTCOND, there appears to be the entries Unknown and Other. We decide to categorise them as missing values as well. This results in 13% of the rows containing missing values. We will decide to fill the missing data with the most common entry for each attribute in the data preparation phase and will investigate if discarding the rows instead improves the model at a later stage.