

Applied Data Science Capstone project

Car accident severity prediction model

Road collisions in Montreal – Canada

Week 1 assignment

1 Problem description

Predicting road accidents is quite uncertain because the relationship between the multiple factors causing an accident is undetermined. If we could predict traffic accidents, corrective actions could be taken to avoid them. The good news is that we have data describing the characteristics of more than 190000 accidents occurred since 2012. The difficulty is that the influencing factors are just too many and complex to analyze them by simple observation. Therefore, the proposed solution is to build a prototype of a machine learning model to predict road accidents and its severity, given at least the weather and the road conditions.

2 Data description and assessment

Montreal Road collision data in csv format is downloaded directly from the web site of the government of Canada. The data set describes accidents occurred within the territory of Montreal since 2012 and recorded by the Police Service.

2.1 Loading the dataframe

The dataset is loaded in a Jupyter Notebook using python and the extension libraries Pandas and Numpy. The dataset has the following general characteristics,

Shape	190552 entries
	68 attributes, named with a tag name.
Documentation	The web site provides detailed documentation describing each tag name and the codes used to register the information.

2.2 The variable to be predicted (Severity)

The dataset includes the attribute GRAVITE. According to the documentation, this attribute holds the severity of the accidents with a descriptive text. There are 5 severity categories, therefore the texts will be replaced by numeric codes, so they can be feed in the predictive model.

	Code
<i>Accident involving minor material damages</i>	1
<i>Accident involving major material damages</i>	2
<i>Accident involving persons with minor injures</i>	3
<i>Accident involving persons with major injures</i>	4
<i>Accident involving fatalities</i>	5

2.3 First attribute: surface condition

The documentation states that the attribute CD_ETAT_SURFC describes the surface condition using codes. It is necessary to replace some of these codes to avoid bias in the predictive model.

	Original Code	Replaced by
<i>Dry surface</i>	11	11
<i>Humid surface</i>	12	12
<i>Water accumulation</i>	13	13
<i>Sand</i>	14	14
<i>Melted ice</i>	15	15
<i>Light snow</i>	16	16
<i>Hardened snow</i>	17	17
<i>Ice</i>	18	18
<i>Mud</i>	19	19
<i>Oil</i>	20	20
<i>Other</i>	99	0
<i>Not registered</i>	NaN	

A new data set is created including the accident severity and the surface condition attributes, using the code descriptors listed above. Then, this new dataset is grouped by categories of accident severity and by doing so, the entries are counted by Surface condition. The results are presented as follows.

Count of entries registered for each characteristic of surface condition, as a function of accident severity

	Dry surface 11	Humid 12	Water 13	Sand 14	Melted ice 15	Snow 16	Hard snow 17	Ice 18	Mud 19	Oil 20	Other 0
<i>Severity</i>											
1	46636	10835	67	81	995	8686	1079	1832	31	3	380
2	42414	12066	125	78	1298	8049	1011	2608	23	9	218
3	26676	7664	63	65	524	2053	281	1021	10	17	41
4	1075	336	2	3	12	58	8	19	1	0	4
5	143	49	0	0	4	8	1	1	0	0	0

It is seen that a good amount of surface condition data has been collected to be able to predict accident severities 1; 2 and 3, but perhaps difficulties might be expected to predict severities 4 and 5, specially under surface conditions different to Dry surface and Humid surface.

2.4 Second attribute: weather condition

According to the documentation, the weather condition is under the tag CD_COND_METEO. Therefore, the dataset also uses codes to describe the weather condition in the time of each accident. Then, these codes were replaced to avoid bias in the predictive model.

	Original Code	Replaced by
<i>Clear sky</i>	11	11
<i>Cloudy</i>	12	12
<i>Mist</i>	13	13
<i>Rain</i>	14	14
<i>Heavy rain</i>	15	15
<i>Windy</i>	16	16
<i>Snow</i>	17	17
<i>Storm</i>	18	18
<i>Ice</i>	19	19
<i>Other</i>	20	0
<i>Not registered</i>	NaN	

This time, a new data set is created including accident severity and the weather condition attribute, using the code descriptors listed above. Then, this new dataset is grouped by categories of accident severity and by doing so, the entries are counted by weather condition. The results are presented as follows.

Count of entries registered for each characteristic of weather condition, as a function of accident severity

	Clear 11	Cloudy 12	Mist 13	Rain 14	Heavy rain 15	Windy 16	Snow 17	Storm 18	Ice 19	Other 10
<i>Severity</i>										
1	49833	10118	108	4303	290	149	4021	501	210	674
2	45577	10278	109	5150	364	156	4772	604	298	340
3	26901	5451	56	3664	361	77	1565	181	97	39
4	1071	208	0	164	16	2	44	7	2	3
5	133	42	2	17	3	4	6	0	0	0

It is seen that a good amount of weather condition data has been collected to be able to predict accident severities 1; 2 and 3, but perhaps difficulties might be expected to predict severities 4 and 5, specially under weather conditions as mist, heavy rain, windy, snow, storm, ice and other conditions.

2.5 Correlation to accident severity

Correlation is an index between -1 to 1, indicating the linear relationship of two dimension. This index is divergent, meaning that the maximum values are 1 and -1, indicating perfect positive or negative linear behavior, whereas 0 means no correlation at all.

The following table shows the correlation index of the Surface condition and weather condition attributes as a function of accident severity.

<i>Severity</i>	Surface condition correlation	Weather condition correlation
1	-0,18	-0,46
2	-0,18	-0,45
3	-0,2	-0,46
4	-0,18	-0,51
5	-0,85	-0,73

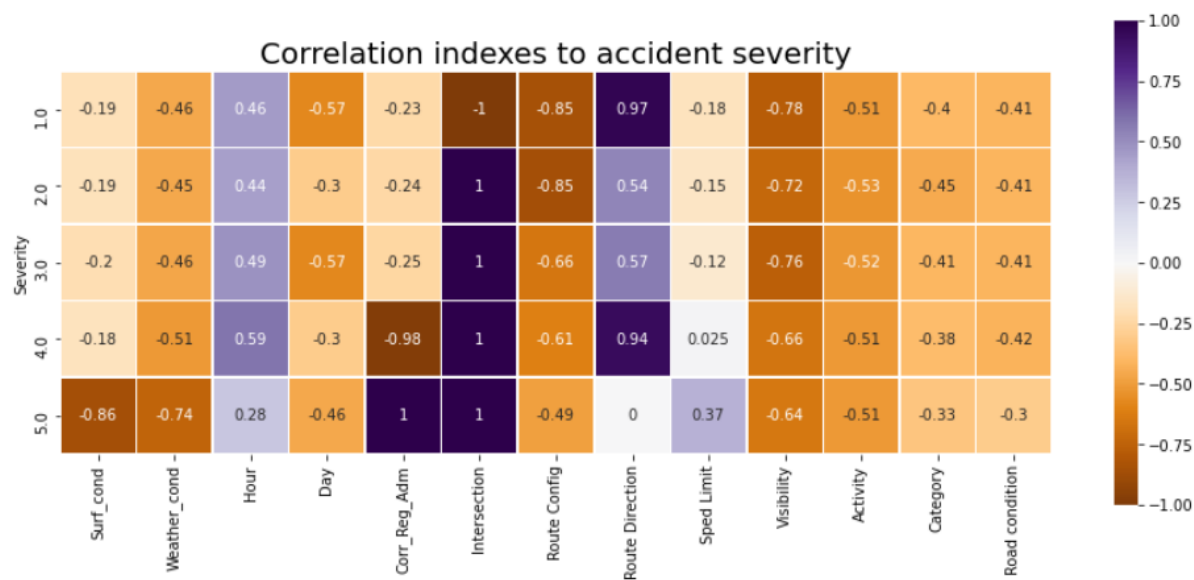
The correlation results show not a good linear relationship of the surface condition to accident severity, whereas the weather condition linearity to accident severity is fair. Perhaps the relationship between these two variables is not linear or they are not the best predictors of severity.

2.6 Assessment of other attributes.

Looking for attributes with better correlation to accident severity, an equivalent procedure to those shown above is now performed to the following attributes,

Tag name	Description
HEURE_ACCDN	Hour of the accident
JR_SEMN_ACCDN	Day of the week
REG_ADM	Region code
TP_REPRR_ACCDN	Proximity to an intersection
CD_CONFIG_ROUTE	Configuration of the road
CD_PNT_CDRNL_ROUTE	Direction of the road
VITESSE_AUTOR	Speed limit
CD_ECLRM	Visibility
CD_ENVRN_ACCDN	Activity of the environment
CD_CATEG_ROUTE	Category of the road
CD_ETAT_CHASS	Road condition

The correlation results, as a function of accident severity are shown in the following heat map.



The heat map above shows with darker colors, attributes with better correlation results to accident severity than the correlation from surface condition and weather condition. Therefore, it is recommended to consider including some of these better predictors to improve the performance of the predicting model.