Applied Data Science Capstone project

Car accident severity prediction model

Road collisions in Montreal – Canada

By, G. Fontecha

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
Accident involving minor material damages	1
Accident involving major material damages	2
Accident involving persons with minor injures	3
Accident involving persons with major injures	4
Accident involving fatalities	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
Dry surface	11	11
Humid surface	12	12
Water accumulation	13	13
Sand	14	14
Melted ice	15	15
Light snow	16	16
Hardened snow	17	17
Ice	18	18
Mud	19	19
Oil	20	20
Other	99	10
Not registered	NaN	Ommited

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

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
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

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
Clear sky	11	11
Cloudy	12	12
Mist	13	13
Rain	14	14
Heavy rain	15	15
Windy	16	16
Snow	17	17
Storm	18	18
Ice	19	19
Other	20	10
Not registered	NaN	Ommited

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

Severity	Clear 11	Cloudy 12	Mist 13	Rain 14	Heavy rain 15	Windy 16	Snow 17	Storm 18	Ice 19	Other 10
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.

Severity	**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; therefore, a classification algorithm should be used instead of a regression one.

2.6 Assessment of other attributes.

Several other attributes are available in the dataset. These additional attributes were not used within the model, but they might be used in the future to improve model predictability. The following table lists these 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_CONFG_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

3 Data preparation

The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation was performed multiple times and it included balancing the labeled data, transformation, and cleaning the dataset.

By this manner, first, the entire dataset was encoded as described in the previous section. Then, the dataset was shuffled and finally split by 70%, used to training, thus reserving 30% for model testing and validation. The dataset was also split into the input data X (surface condition and weather condition) and the predicting result y (accident severity).

The predicting result has 5 categories of accident severity; therefore, 5 copies of the data was used for each accident category. By this way, each category in each copy may be set in 1 and the other categories may be set in 0.

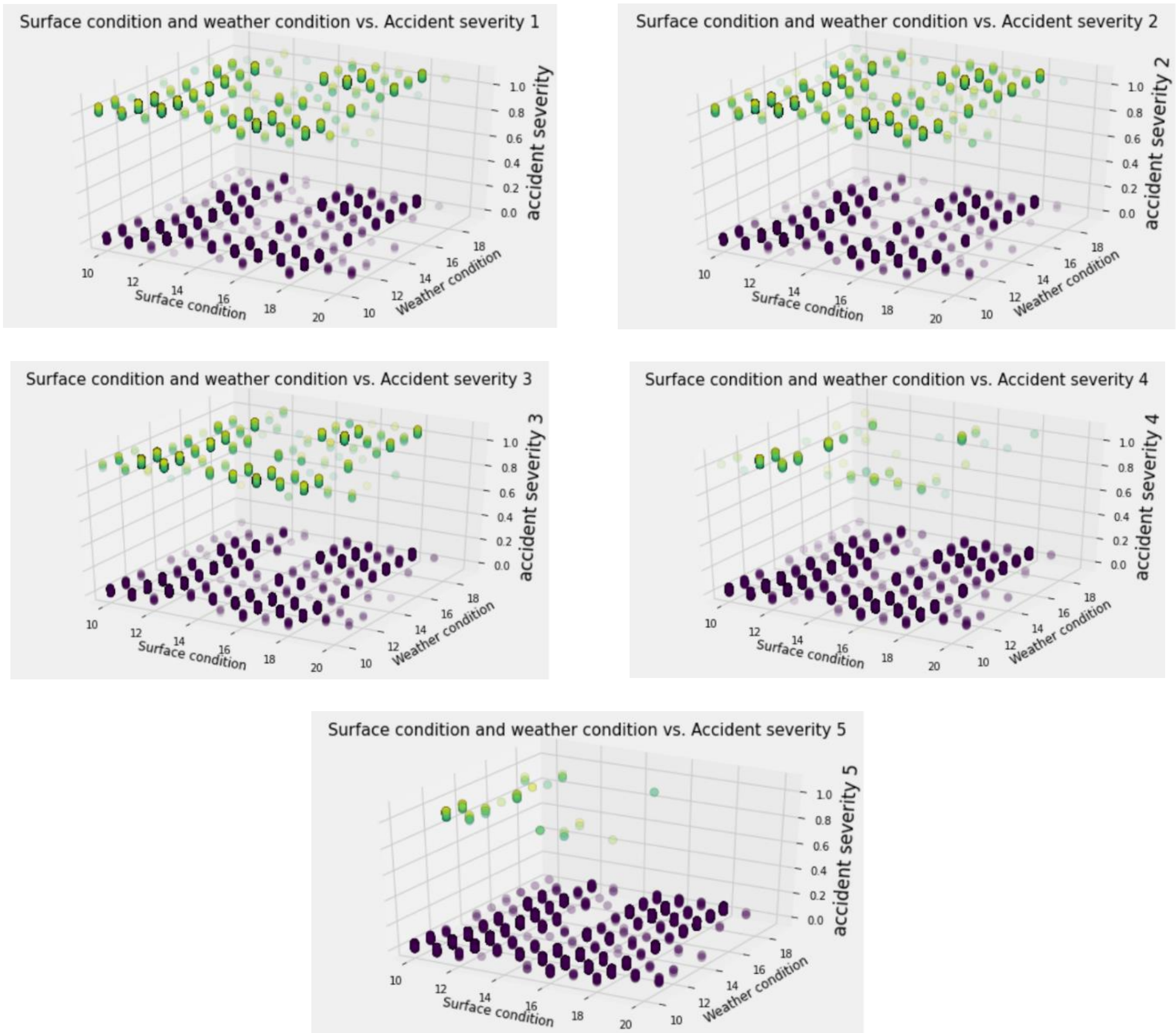

Figure 1 scatter plots showing the patterns for each accident severity category.

The scatter plots in Figure 1 show the accident severity for each accident severity category as a function of surface condition and weather condition. In those figures, it is possible to recognize some patterns which may be more distinguishable by a classification algorithm.

4 Modeling and model assessment

In this phase, various algorithms and methods are selected and applied to build the model including supervised machine learning techniques.

4.1 K-nearest Neighbors

A K-Nearest Neighbors model was ran, using *k* factors between 3 and 8 but its prediction capacity is quite poor, apparently because in this application there are conditions triggering both the occurrence of an accident or not an accident at all.

4.2 Decision tree, Support Vector Machine and Logistic regression

By the contrary of the K-Nearest Neighbors model, these tree models successfully predicted the occurrence of an accident at every category group. Two key elements were important to this success; first, the data must be very well shuffled; second, splitting the data into accident categories allow highlight the differences between each accident category and the other categories.

Figure 2 show the confusion matrixes using a support vector machine model for each accident severity. However, the same results were obtained using whether a decision trees model or a logistic regression model.

In terms of computing time, both the decision trees model and the logistic regression model are done in about 1 second, and 2 seconds in the case of the support vector machine model. However, parallel computing was configured to use all the 6 cores of an intel i5-9000 CPU @ 3.0 GHz.

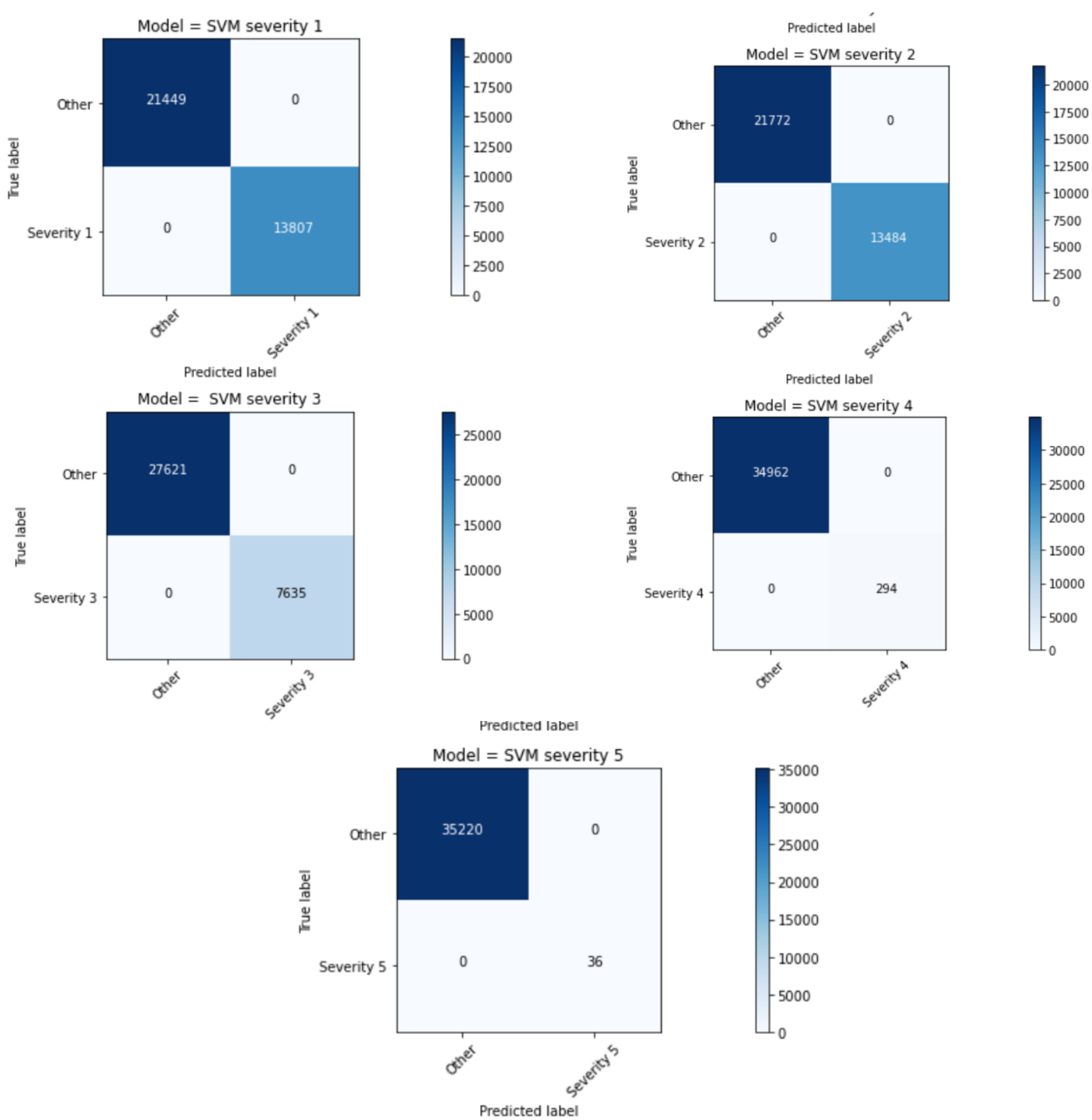

Figure 2 Confusion matrix for each accident severity category.

5 Deployment

The code is attached in a github, under a jupyter notebook file. Also attached are the python file functions called for parallel computing. The code runs the models for every dataframe categories, said accidents classified as severity 1, severity 2, severity 3, severity 4 and severity 5.