Applied Data Science Capstone Project

Ambarish Ambuj

Introduction

The severity code of the accident is typically set such that it represents the extent of damage caused by the accident. In an environment of limited resources, focusing more resources on preventing high severity accidents is one of the solutions to minimize the amount of damage with given resources. However, to do that, an understanding of the factors that affect the severity of the accident and the extent to which they affect the severity, is essential. Hence, with the given data about accident severity and some related parameters, this project tries to come up with a model to predict the impact of some key parameters such as accident location type, collission type, weather condition, road condition, lighting condition, number of persons involved in the collision etc. on the severity of the accident. The output of this model can provide policy inputs to the government to take specific actions to mitigate the causes that impact the accident severity the most.

Data Description

The base data is taken from the example dataset provided in the course at the link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

We will first import the data as a dataframe to get a glimpse of the data.

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	NaN	NaN
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	NaN	6354039.0
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	NaN	4323031.0
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	NaN	NaN
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	NaN	4028032.0

So, there are 194673 observations of incidents. There are 38 columns in the original dataset but as is evident from a preview of first 5 rows of the data, a column called 'SeverityCode' is repeated. So, there are 37 attributes for 194673 incidents. However, going back to our problem definition, not all 37 attributes are of our interest. We are only interested in exploring the impact of certain mitigable attributes on severity of the accident. So, based on the primary theoretical understanding, we select 'SEVERITYCODE' as the dependent variable and following variables as dependent variable:

1) 'ADDRTYPE': A catagorical variable representing the type of location where incident took place. It may take the values of 'Intersection', 'Block' etc.
2) 'COLLISIONTYPE': A categorical variable indicating the type of collision such as head-on, angle etc.
3) 'PERSONCOUNT': An integer representing number of persons involved in the collision.

4) 'PEDCOUNT': An integer representing number of pedestrians involved in the collision.
5) 'PEDCYLCOUNT': An integer representing the number of bicycles involved in the collision.
6) 'VEHCOUNT': An integer representing the number of vehicles involved in the collision.
7) 'WEATHER': A categorical variable describing whether the weather was cloudy or rainy etc. at the time of collision
8) 'ROADCOND': A categorical variable describing condition of the road i.e. dry or wet
9) 'LIGHTCOND': A categorical variable describing the lighting condition at the time of collision.

So, let's extract the one target and 9 predictor variables from the dataframe and store it in a new dataframe.

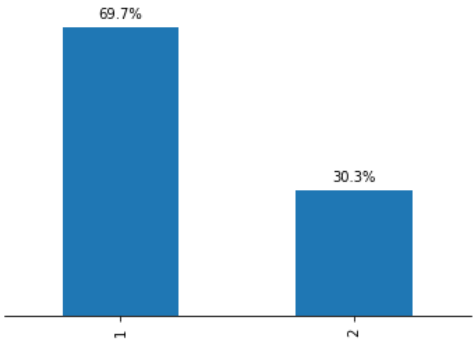
	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	WEATHER	ROADCOND	LIGHTCOND
0	2	Intersection	Angles	2	0	0	2	Overcast	Wet	Daylight
1	1	Block	Sideswipe	2	0	0	2	Raining	Wet	Dark - Street Lights On
2	1	Block	Parked Car	4	0	0	3	Overcast	Dry	Daylight
3	1	Block	Other	3	0	0	3	Clear	Dry	Daylight
4	2	Intersection	Angles	2	0	0	2	Raining	Wet	Daylight

As we have a large number of data available with us, the best treatment of missing values is to drop them so that we don't have to guess the missing values and thereby affect the output. So, we will drop any rows with missing values. So, some of the rows were dropped and now we have 187504 observations with us to train, validate and test the model.

Exploratory Data Analysis

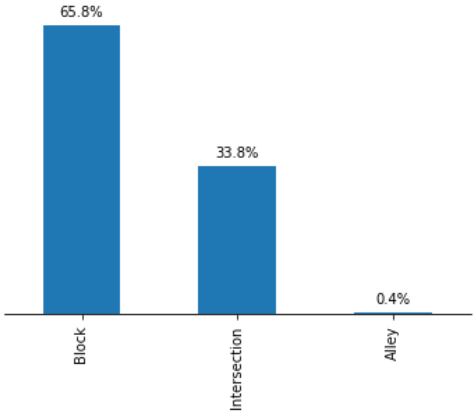
We will first explore the data and try to observe some patterns within the data which may further help in our analysis exercise.

1) Distribution of SEVERITYCODE:



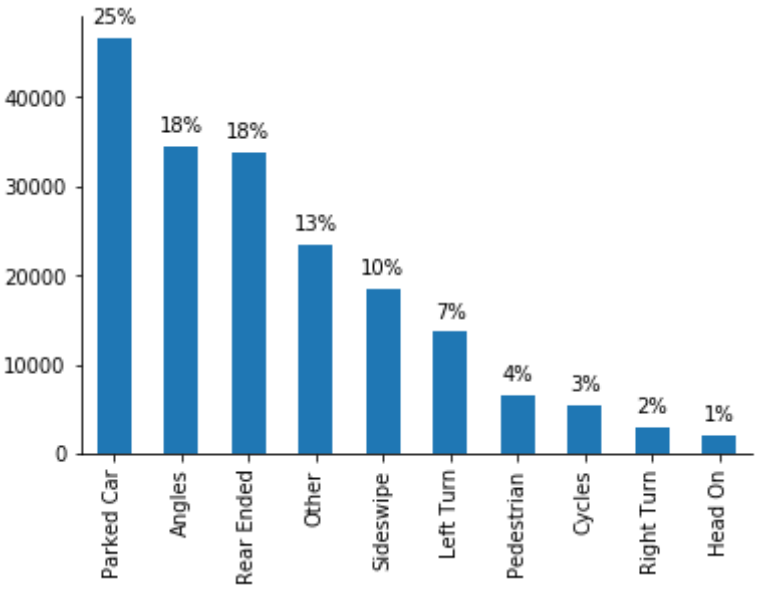
As is evident from the bar chart, 69.7% of all accidents have been of severity 1 i.e. only property damage whereas remaining 30.3% accidents resulted in some human injury as well.
This is on expected lines as we expect more accidents of less severity.

2) Location of accidents:



It indicates that about two third of all accidents took place in blocks whereas about one third took place at intersections. Alleys, understandably, contributed negligible proportion of accidents.

3) Distribution of accidents by Collision type:

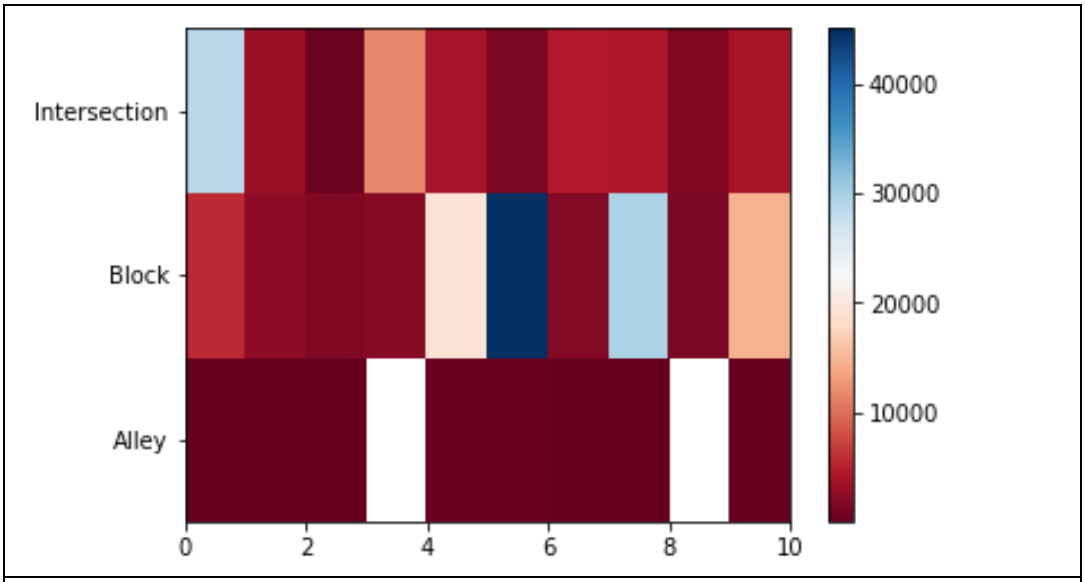


This bar chart provides some very interesting insights. It indicates that about a quarter of all accidents involved a parked car. It is likely that these incidents are mostly happening in blocks rather than intersection. This may be one plausible reason why blocks have more accidents than intersections.
This also provides an interesting policy question to address and regulate the parking in blocks to avoid these accidents.
'Angles', 'Rear Ended' and 'Sideswipe' are other prominent types of accidents.

4) Location of Accident & Collision Type:

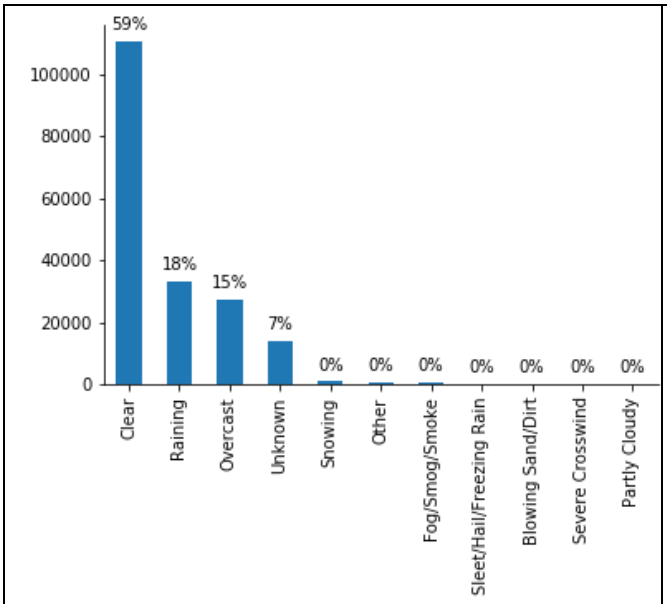
Looking at the accident location and collision type in isolation itself has given us significant insights. However, looking at them together may provide us further insights. So, we create a pivot table and a heatmap to understand it better.

COLLISIONTYPE	Angles	Cycles	Head On	Left Turn	Other	Parked Car	Pedestrian	Rear Ended	Right Turn	Sideswipe
ADDRTYPE										
Alley	57	8	4	NaN	284	325	37	11	NaN	16
Block	5653	2298	1567	2114	19416	45057	1856	29595	1226	14533
Intersection	28845	3093	440	11545	3740	1297	4696	4188	1710	3893



The pivot table and the heatmap confirm our previous hypothesis about parked car contributing a significant proportion of accidents reported in blocks. It is followed by the 'angle collision at intersections', 'sideswipe collision in blocks' and 'left turn collision at intersections'. If these four issues could be addressed systematically, about 60% of the accidents can be avoided.

5) Distribution of weather type for accidents:

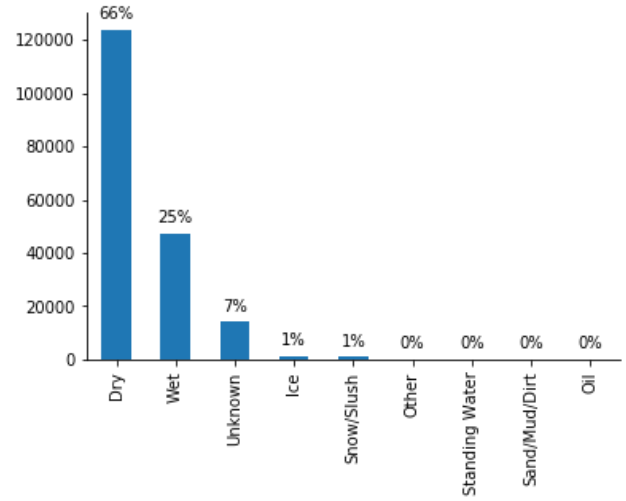


SEVERITYCODE	1	2
WEATHER		
Blowing Sand/Dirt	36	13
Clear	74775	35718
Fog/Smog/Smoke	377	186
Other	676	114
Overcast	18834	8711
Partly Cloudy	2	3
Raining	21835	11134
Severe Crosswind	18	7
Sleet/Hail/Freezing Rain	85	27
Snowing	729	167
Unknown	13267	790

This chart indicates that 59% of the accidents took place on clear days, 18% on rainy days and 15% on overcast days.

On clear, rainy as well as overcast days, the ratio of severity 1 and severity 2 accidents appear to be similar indicating that the weather may not have significant effect on severity of the accidents.

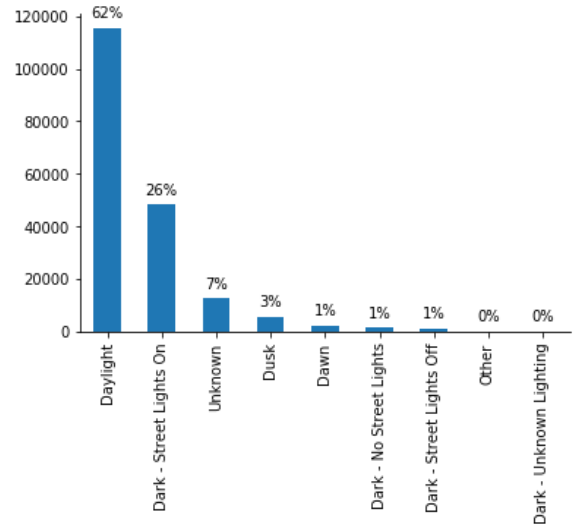
6) Distribution of Road Condition for accidents:



SEVERITYCODE	1	2
ROADCOND		
Dry	83832	39898
Ice	923	269
Oil	40	24
Other	82	42
Sand/Mud/Dirt	51	22
Snow/Slush	827	165
Standing Water	82	29
Unknown	13276	729
Wet	31521	15692

The proportion of wet among severity 2 accidents is slightly higher than the proportion of wet among all accidents. This may indicate a role of wet roads in increasing the severity of the accident which can be further evaluated using machine learning models.

7) Distribution of Light Condition for accidents:



SEVERITYCODE	1	2
LIGHTCOND		
Dark - No Street Lights	1191	334
Dark - Street Lights Off	869	315
Dark - Street Lights On	33816	14417
Dark - Unknown Lighting	7	4
Dawn	1667	823
Daylight	76995	38400
Dusk	3906	1936
Other	175	52
Unknown	12008	589

Lights do not seem to be a problem as most of the accidents occurred in daylight or with street lights on.	The data does not point to any obvious relation between light condition and accident severity.

Methodology

To explore the predictability of accident severity based on the selected explanatory variables, we will train a classifier model on the data.
We will follow the following steps to arrive at a classifier model:

1. First of all, we will create the dummy variables for the categorical variables.
2. Next, we will split the available dataset into training, cross-validation and test data sets. Training data will be used to train the models, cross validation data will be used to fine-tune the model by adjusting certain parameters, and the test data will be used to evaluate the performance of the models.
3. We will fit logistic regression model to the data and evaluate the accuracy.

Results

Data Pre-processing

First we create dummy variables for all categorical variables and then transform all variables to zero mean and unit standard deviation.

After transformation, we split the dataset into 60:20:20 i.e. 60% for training, 20% for cross validation and 20% for testing.

Logistic Regression

We train the logistic regression for C ranging from 0.01 to 1 on the training data. We use cross validation data for finding the best C. C=0.3 gives the highest accuracy for cross validation data. So, we select c=0.3 for the final model. We use test data to evaluate the accuracy of final model. The F1 score is 0.718 , the Jaccard Similarity Score is 0.757 and the log loss score is 0.481. This indicates that the model has performed reasonably well in predicting the severity of the accidents based on the variables selected.

Discussion

The predictor variables and the coefficients of this model are in the table on next page.

It is worth noting that the model considers severity 1 as '0' and severity 2 as '1' case. So, the variables with positive coefficients are more likely to be correlated to severity 2 accidents whereas the variables with negative coefficients are likely correlated to severity 1 accidents.

It follows from the table above that 'Parked car accidents' and 'Sideswipe accidents' have strong correlation with severity 1 accidents. This confirms our hypothesis based on exploratory data analysis.

The count of persons, pedestrians, cycles, and vehicles involved in accidents are all positively correlated with the severity 2 accidents i.e. more the number of persons or vehicles involved in the accident, more are the chances of it being a severity 2 accident.

However, the number of persons and number of vehicles have much lower coefficient than number of pedestrians and number of cycles. So, the accidents involving pedestrians and cyclists are more likely to be severity 2 accident than the accidents involving vehicles.

'Overcast' and 'raining' weather conditions have positive coefficients indicating they contribute to severity 2 accidents. Positive coefficient for 'wet' road condition further validates this point.

Among light conditions, 'Daylight' and 'Street Lights on' have highest positive coefficients, indicating that bad light had no significant impact in increasing the severity of the accident.

Conclusion

The model provides significant insights into the factors that determined the severity of accidents. The insights also provide inputs for the policy-makers to avoid high severity accidents.

However, data on more factors and data about higher severity accidents would provide more insights.

	Predictors	Coefficient
0	PERSONCOUNT	0.196347
1	PEDCOUNT	0.503117
2	PEDCYLCOUNT	0.557418
3	VEHCOUNT	0.177225
4	Block	0.384537
5	Intersection	0.464946
6	Angles	0.069204
7	Cycles	-0.034651
8	Head On	0.049831
9	Left Turn	0.044486
10	Parked Car	-0.807155
11	Pedestrian	0.139716
12	Rear Ended	0.153727
13	Right Turn	-0.084374
14	Sideswipe	-0.337537
15	Blowing Sand/Dirt	0.002608
16	Clear	0.172218
17	Fog/Smog/Smoke	0.017697
18	Overcast	0.105056
19	Partly Cloudy	0.014706
20	Raining	0.104690
21	Severe Crosswind	0.002489
22	Sleet/Hail/Freezing Rain	-0.011673
23	Snowing	0.015935
24	Dry	0.303440
25	Ice	0.042644
26	Oil	0.024158
27	Sand/Mud/Dirt	0.019980
28	Snow/Slush	0.020552
29	Standing Water	0.012825
30	Wet	0.276915
31	Dark - No Street Lights	0.038395
32	Dark - Street Lights Off	0.047863
33	Dark - Street Lights On	0.264161
34	Dark - Unknown Lighting	-0.002439
35	Dawn	0.073487
36	Daylight	0.289292
37	Dusk	0.114966