Predicting the Severity of the Seattle Road Accidents

Nikhil Panda

6th September, 2020

Business Problem/Interests

- The data we have is about accidents in Seattle city. With the data we have, we can predict the severity of the accident. So my attempt here or problem I choose here is to derive the severity of the accident. Generally in accidents, the primary things we would verify are road condition and visibility on the road and sometimes the weather plays a major role especially during rains and winter. Besides these speeding is the key indicator to judge the accident's cause. As all this information is there in that Seattle accident's data.

- As we are predicting the severity of the accident here, the primary audience will be the traffic police to do the primary analysis at offsite and take further steps based on the outcome. Besides them, insurance stakeholders can use this to predict the likely outcome if their policyholders had involved in this as an initial investigation report offsite. And a small set of the group would be the people who can use to know the likelihood of an unexpected journey based on factors like weather, road & lighting condition.

Data acquisition and cleaning

- The data in this data source is provided by SPD and recorded by Traffic Records. This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.
- We had circa 193K+ records for accidents during the time span starting from 2004 to the recent past with 37 Independent variables and outcome or labelled data will severity of the accident. Severity has been categorized into 5 categories by SDOT.

  3—fatality

  2b—serious injury

  2—injury

  1—prop damage

  0—unknown
- Out of all records, only circa 30% is having the predicted severity '1' records and the rest were severity 2. So have balanced the number of records for both the outcomes to prevent the bias towards one outcome as majority records can pull off many variations in prediction.

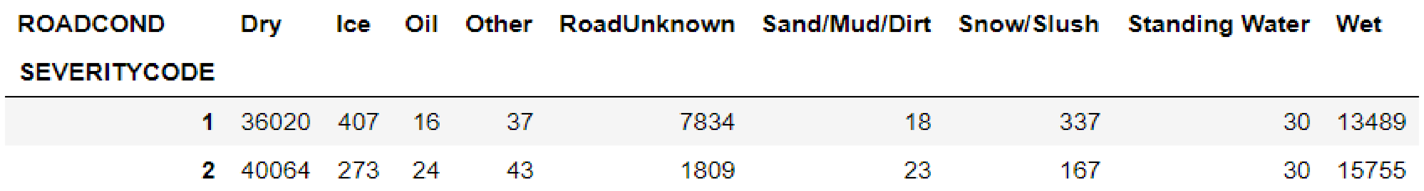
Feature Selection and Data preparation

- So have selected all the key factors present in the data for prediction which are
  1. Weather condition.
  2. Road condition.
  3. Lighting condition.
  4. Collision Address type
  5. Speeding flag

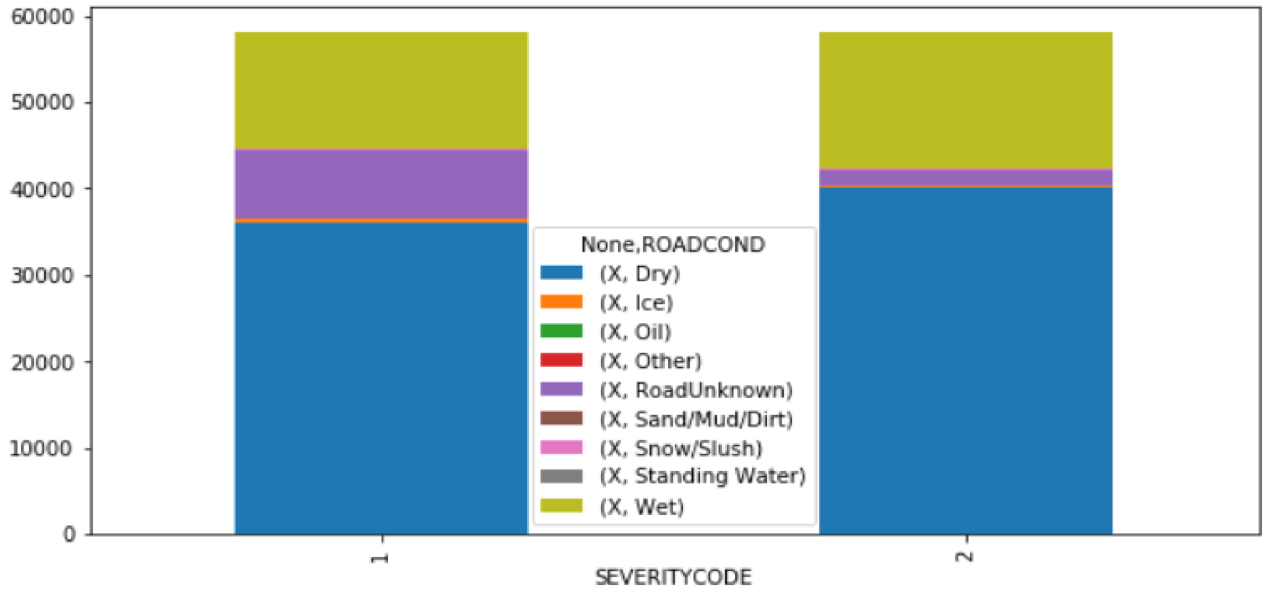
- After selecting the data, I have seen some missing values and have updated them as follows.
- For weather, road condition, light condition & Address type data, have replaced the missing values with the unknown as already a certain number of records were having that values. A majority number of records are not having any data in the speeding flag, so assumed speeding was not there and replaced with no value as speeding not recorded.
- After doing this sanitation, I have seen data in good form for normalizing and looked perfect as categorical values.

Data exploration (contd..)

- Relation between Road condition with severity code of the accident.

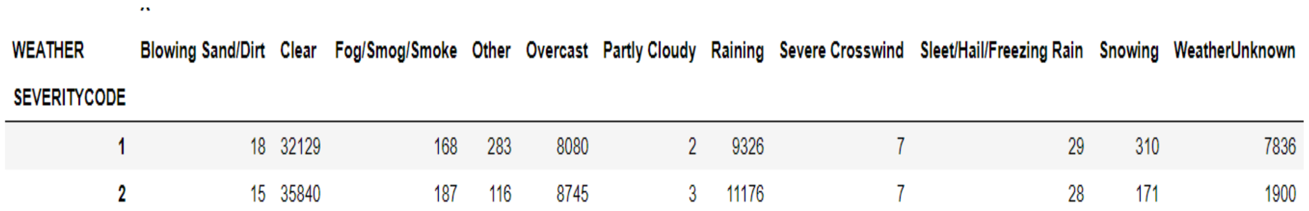


ROADCOND	Dry	Ice	Oil	Other	RoadUnknown	Sand/Mud/Dirt	Snow/Slush	Standing Water	Wet
SEVERITYCODE									
1	36020	407	16	37	7834	18	337	30	13489
2	40064	273	24	43	1809	23	167	30	15755

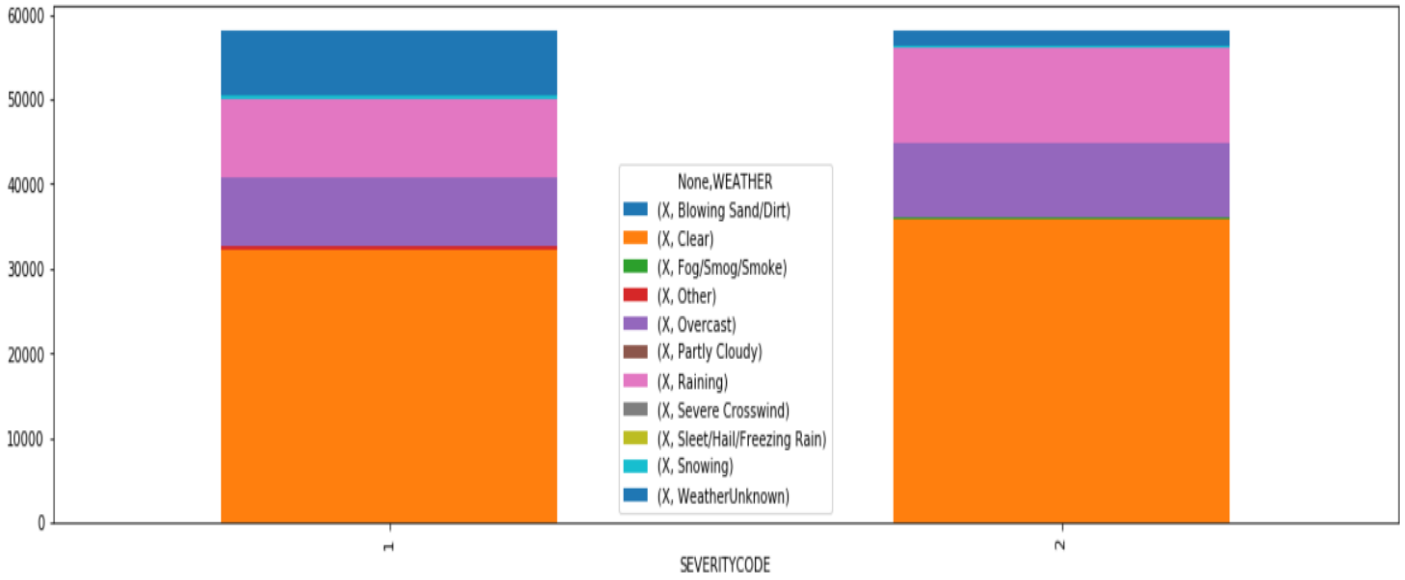


Data exploration (contd..)

- Relation between Weather condition with severity code of the accident.

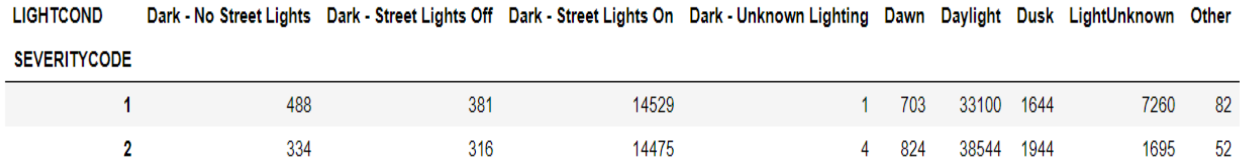


WEATHER	Blowing Sand/Dirt	Clear	Fog/Smog/Smoke	Other	Overcast	Partly Cloudy	Raining	Severe Crosswind	Sleet/Hail/Freezing Rain	Snowing	WeatherUnknown
SEVERITYCODE											
1	18	32129	168	283	8080	2	9326	7	29	310	7836
2	15	35840	187	116	8745	3	11176	7	28	171	1900

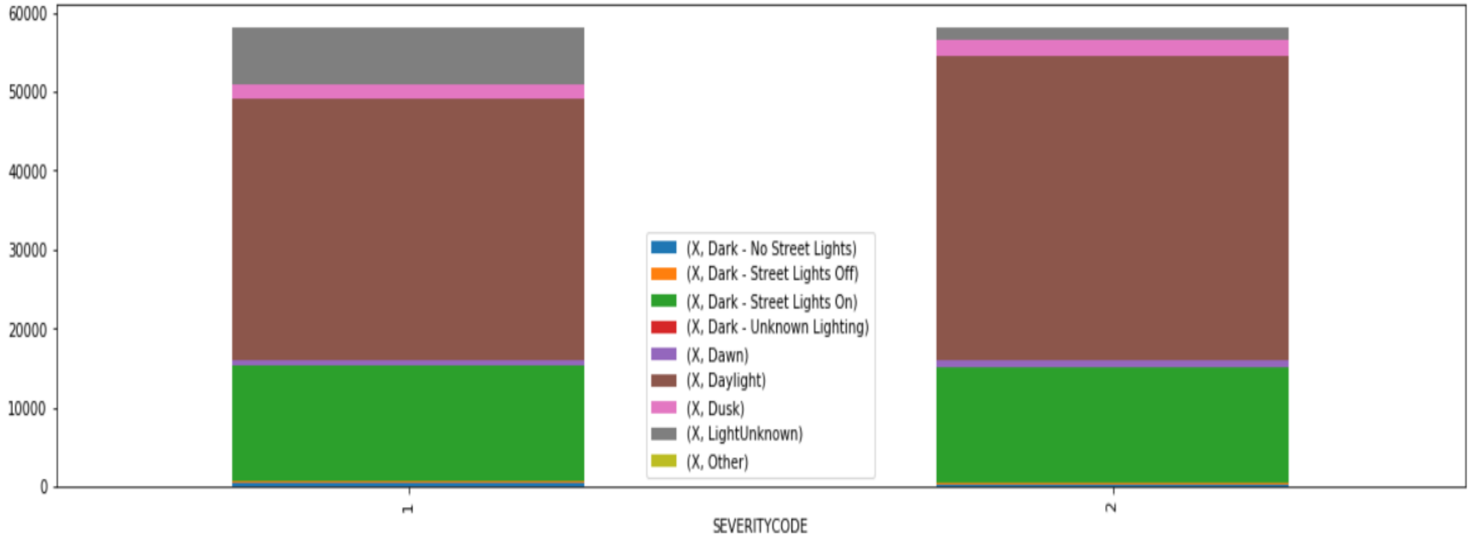


Data exploration (contd..)

- Relation between Lighting condition with severity code of the accident.

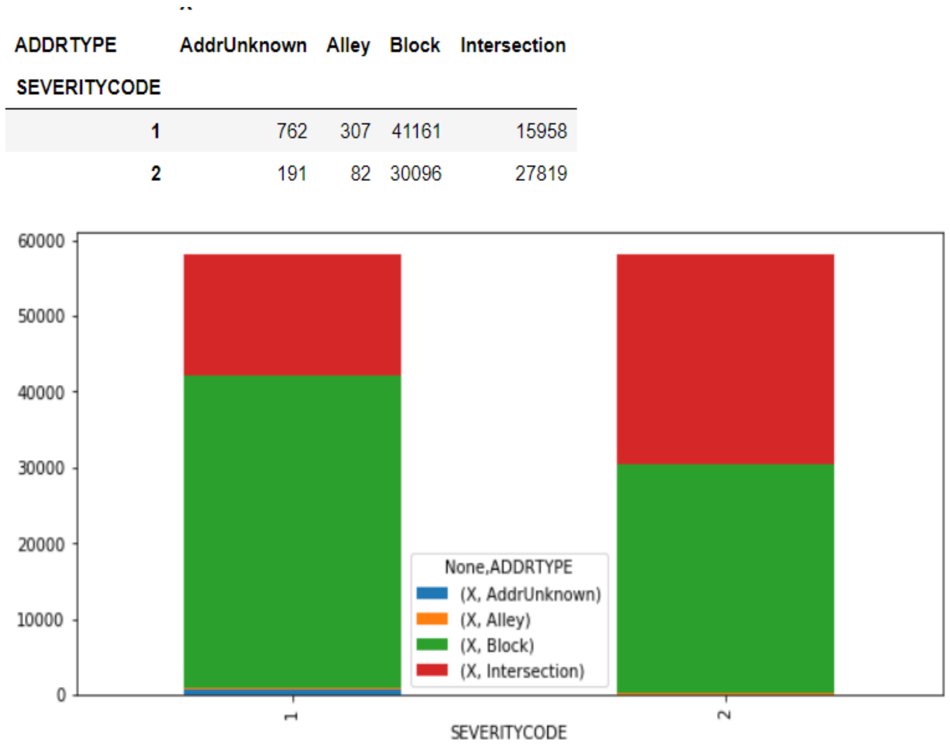


LIGHTCOND SEVERITYCODE	Dark - No Street Lights	Dark - Street Lights Off	Dark - Street Lights On	Dark - Unknown Lighting	Dawn	Daylight	Dusk	LightUnknown	Other
1	488	381	14529	1	703	33100	1644	7260	82
2	334	316	14475	4	824	38544	1944	1695	52

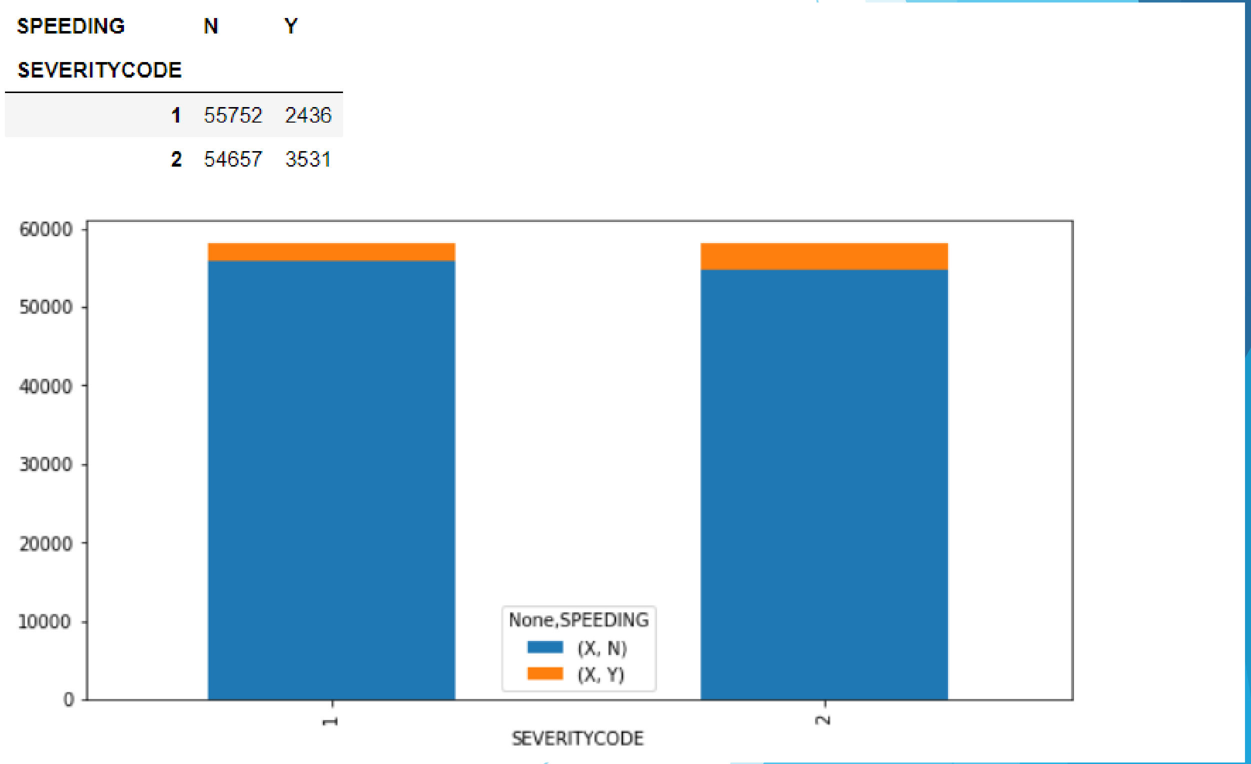


Data exploration (contd..)

- Relation between Collision address type condition with severity code of the accident and Speeding Index with Severity code.



ADDRTYPE	AddrUnknown	Alley	Block	Intersection
SEVERITYCODE				
1	762	307	41161	15958
2	191	82	30096	27819



SPEEDING	N	Y
SEVERITYCODE		
1	55752	2436
2	54657	3531

Methodology/Predicting Model

- As the outcome of the model will be a classified class, so we will use the classification models here for building the model. We can use any of these 3 models to build the data.
  - K Nearest Neighbor(KNN)
  - Decision Tree
  - Support Vector mission (SVM)

Metrics	KNN (K = 18)	Decision Tree	SVM
Jaccard Index	0.530	0.608	0.530
F1 Score	0.529	0.605	0.529
Precision	0.531	0.608	0.530
Recall	0.530	0.612	0.531

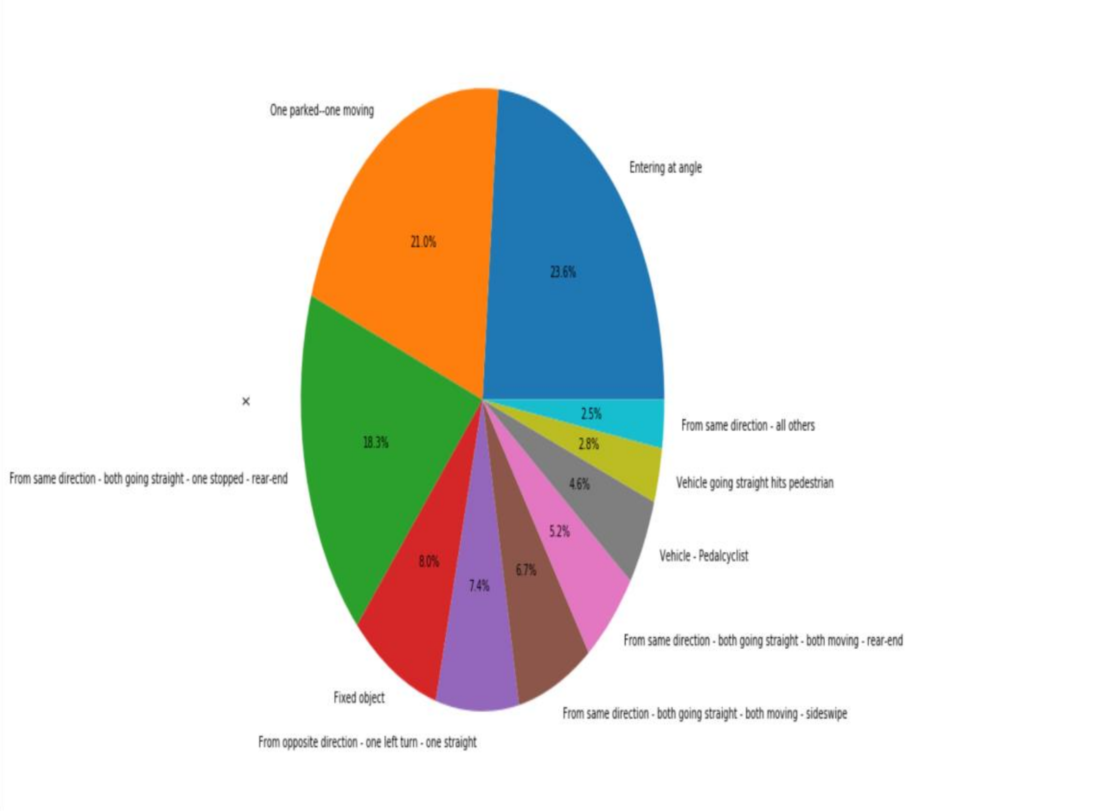
After evaluating the scores for all the models we trained, Decision tree looks better than remaining with accuracy rate around 60%, model accuracy can be increased with help of more data as this data is not having all the possible outcomes.

Conclusion: (contd..)

- In this case study, I have analyzed the Seattle accidents data to predict the severity of the accident using the 5 features weather condition, road condition, lighting condition, collision address type & speeding index.
- The results highlighted the unexpected insight as most accidents were happening when all the things were perfect like most accidents occurred during day times if we thought that lighting may be reason.
- Most accidents occurred during the clear sky if we think rains, fog or snow might cause the accident.
- Likewise most accidents happened on the dry roads if we think road’s behavior may be the reason.
- And in context of speeding, only 10% of the selected data are having speeding Index is Y, so speeding is not the complete reason. Many accidents happened across the intersections & blocks.

Conclusion: (contd..)

- But the one important finding I have noticed from the data which is not a feature variable to predict the severity, the field ‘ST_COLDESC’ which is post investigation outcome by SDOT to categorize the collision.
- The pie chart depicts the share of the most collision types in the data, and the shocking finding is that collisions happened mostly with parked vehicle, with vehicle going in same direction and vehicles entering in angels.



The conclusion from this finding is to manage the vehicles going in same direction to follow proper distancing between the vehicles going in the same direction and following the proper lines to avoid slipping into parking lanes.

Riders should make use of side signals mostly to avoid the accidents between the vehicles going in same direction or with parked cars.

Future Directions:

- Most of the data is not fitted into training data due to lack of all possible scenarios, it would be good to include data with all possible outcomes instead of only Severity code 1 & 2.
- We can avoid training this data into SVM model as it is taking lot of time and data is not very highly dimensional too.

Thank you