Severity Prediction Model - Data Approach to Risk Control

Aditya WB

September 28, 2020

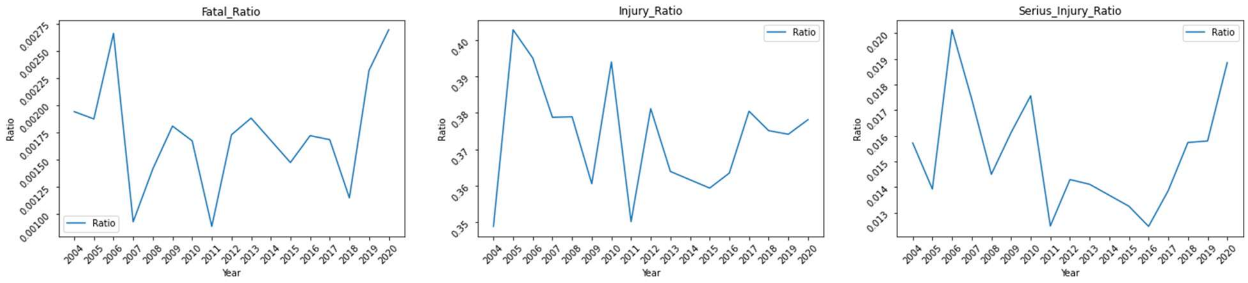
Table of Content

1.0 Introduction

1.1 Background

In the last 5 years Seattle have seen increase in fatal accidents and accidents with serious Injury. The Charts below show traffic accidents since 2014 and 2015 have increased in the number of fatalities and serious injury despite improvement that the city have from 2010. This means in the last 5 years the road is not as safe as before. If not intervene, the community will carry more and more burden from the medical cost and from loss of income to families due to disabilities and death. Insurance company may also be forced to adjust the automobile liability premium up for customer living or commuting in Seattle in which it is a lose-lose situation for both Insurance Industry and community as a whole.

Chart Below taken from the Seattle city GIS Collision Data



1.2 Problem

To build a model that can provide indication whether a collision accident in certain road, weather, or driver condition will involve bodily injury or worse death. Data used to build the model should include i.e. the weather condition during the accident, light condition, how the collision happened and so on.

The model can then be used by city government to develop regulation or to build infrastructure to prevent serious injuries or fatalities.

1.3 Interest

In General the community as a whole will benefit from the exercise however, City Government can get more tangible benefit since the model could help to build better regulation, build more collision prevention infrastructure in area where it needs the most, or to plan and manage better traffic management and emergency service.

2.0 Data Acquisition & Cleaning

2.1 Data Source

For this exercise, we will use Collision record from Seattle city GIS. The Documentation for the Data can be found here and the link to the Data Source Is here

The Data contains all collision accidents since 2003 and it clearly describe what kind of collision, severity of the collision, what was the weather, road condition during the accident, how many vehicles involves, how many person involves, degree of injury, etc.. However not all feature of the data are useful for to provide indication to an accident.

The next Section, the Preliminary Data cleaning, we will focus on identifying feature that can be used to build model.

2.2 Preliminary Data Cleansing

Upon Examination we have found out that the Seattle GIS Data is unique that all rows indicate single collision accident without duplicates.

One level Features will be omitted since they won't contribute to the model, EXCEPTRSNCODE is also omitted despite having 2 level due to missing information. Other features below such as 'REPORTNO', 'INCKEY','OBJECTID', 'COLDETKEY','SDOTCOLNUM' are omitted because they are the Unique record identifier which don't carry information on how/ when/ what collision.

Features below are not included because they represent the magnitude of the incident and not the probability of the severe accident (accident where people died).

- PERSONCOUNT (The total number of people involved in the collision).
- PEDCOUNT (The number of pedestrians involved in the collision).
- PEDCYLCOUNT (The number of bicycles involved in the collision).
- VEHCOUNT (The number of vehicles involved in the collision).
- INJURIES (The number of total injuries in the collision).
- SERIOUSINJURIES (The number of serious injuries in the collision).
- FATALITIES (The number of fatalities in the collision)

Features below also not included since they act as description to an existing codes:

- SEVERITYDESC description of SEVERITYCODE
- SDOT_COLDESC description of SDOT_COLCODE
- ST_COLDESC description of ST_COLCODE

Features below are not included since they act as pointer to the collision location but not the feature of the location:

- X
- Y
- INTKEY
- LOCATION
- SEGLANEKEY
- CROSSWALKKEY

2.3 Feature Selection

After understanding the meaning of each feature, 10 features will be used as independent variables to SEVERITYCODE. They Are:

- ADDRTYPE :
    - Simple classification of the accident location 'Alley','Block','Intersection'.
- COLLISIONTYPE :
    - Describe how the collision happen such as collision at an angle, involving, cyclist, head on collision, etc.
- JUNCTIONTYPE :
    - Describe the nature of the junction such as at intersection but not related to intersection, at mid-block but intersection related, at ramp junction, etc.
- SDOT_COLCODE :
    - This is set of code define by Seattle's Department of Transportation that clearly define the nature of the collision. For example, it has specific code for collision where a motor vehicle hit a cyclist head on, or whether a motor vehicle rear-ended another vehicle. This feature will be instrumental in determining bodily injury in a collision.
- ST_COLCODE :
    - Similar to SDOT_COLCODE, this collision code prescribed by State Government.
- UNDERINFL :
    - This is Boolean feature that describe whether the driver is under influence (alcohol or any other substance) while driving.
- ROADCOND :
    - This feature describe the road condition during the accident whether it was wet, has shown on it, dry, has sand or dirt etc.
- WEATHER :
    - Describe the weather during the accident, whether it was raining, has smoke or smog, overcast, etc.
- LIGHTCOND :
    - Describe the light condition during the accident, whether it was Dark with street light, completely dark, day light, etc.
- HITPARKEDCAR :
    - This is Boolean feature that describe whether the accident involves hitting a parked car.
- INCDTTM :
    - This the incident date information, for the model we would like to see how weekdays or weekend affects the severity of the collision.

For the target label we will use:

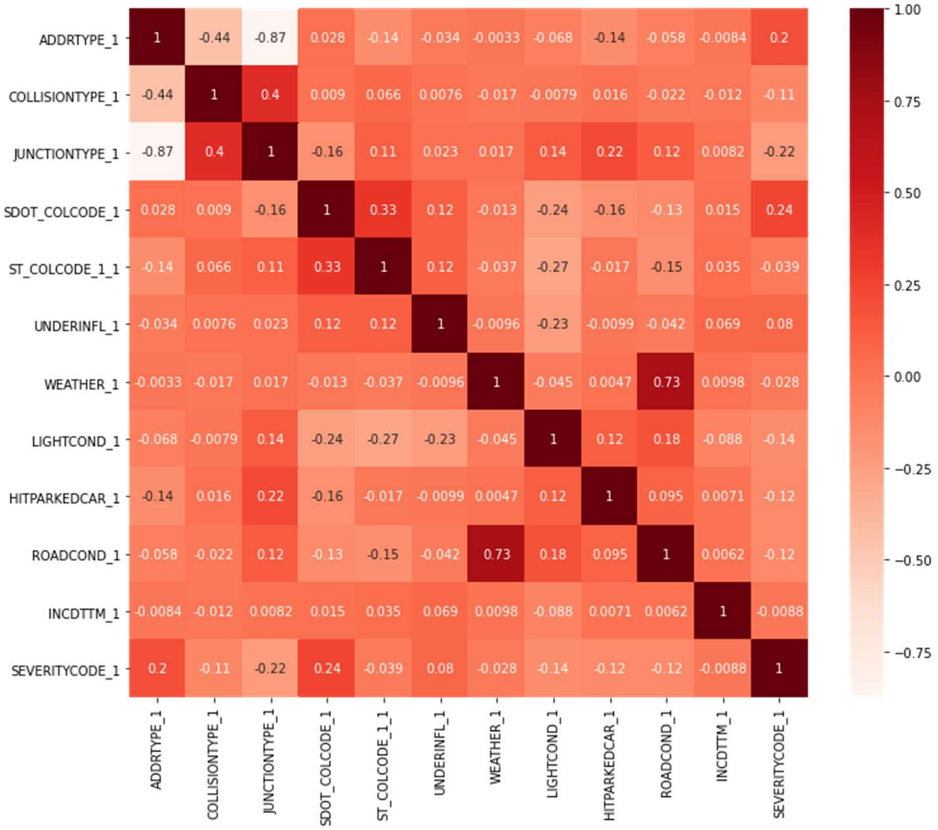
- SEVERITYCODE but transformed to simplify the non-Bodily injury code where all non-bodily injury and death are mapped to '0'.

3.0 Exploratory Analysis

3.1 Correlation Review

The selected features will be reviewed to see whether there are feature that have very little correlation with difference in Severity or features that may be highly correlated with Each other that they may be redundant. For the correlation review, instead of the original feature, the features are numerical coded to allow correlation review.

Pearson Correlation Heat Map on Initial selection



The analysis indicates:

Relatively Low correlation between 'SEVERITYCODE' and the features, the highest correlation score is 24% which is quite low. Also, Heat map examination that there are 2 groups of features that are more inter-correlated:

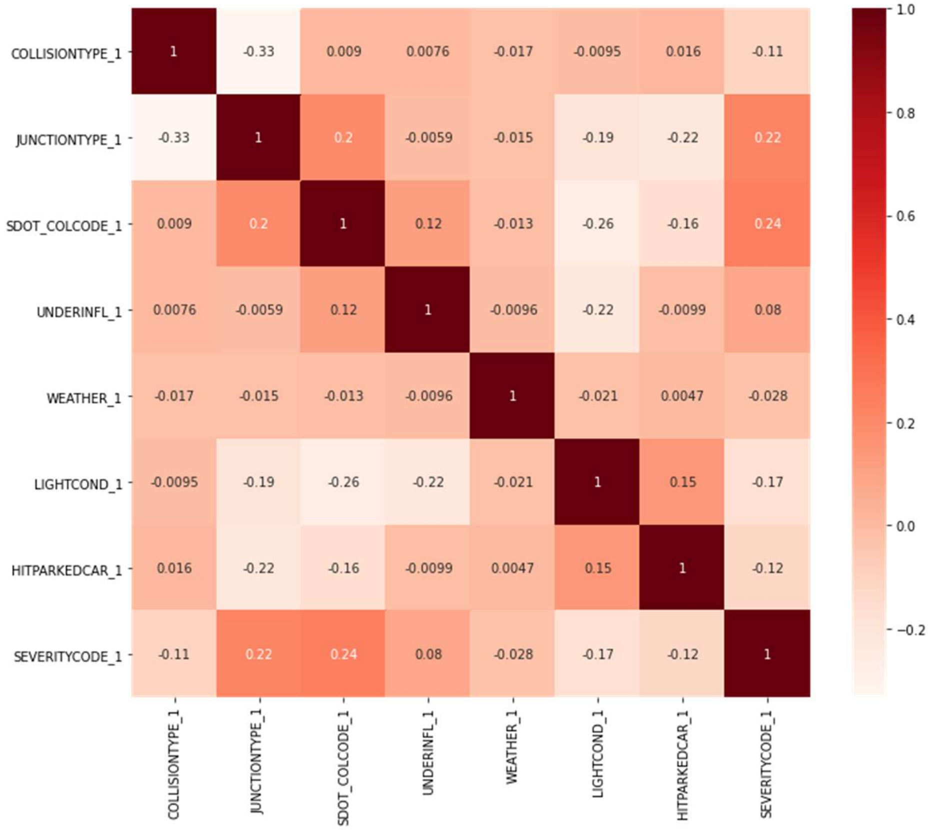
Address Type, Collision Type & Junction Type
Weather, Light Condition & Road Condition
Incident Date (day of the week) feature is proven to be very un-instrumental with less than 1% correlation. Thus Incident Date feature will be dropped from the exercise

Furthermore, we found out that the label coding is not fully aligned yet, for example, for Address type, the max value refer to Intersection but in Junction type, the max value does not refer to intersection type junction. The label will be remapped to allow more consistent correlation

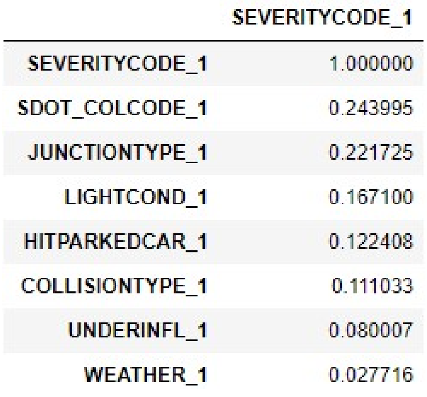
The Correlation analysis to be re-run using the re-labeled JUNCTIONTYPE_1 and with these features removed:

- ADDRTYPE_1 due to high inter-correlation with JUNCTIONTYPE_1
- ROADCOND_1 due to high inter-correlation with WEATHER_1
- INCDTTM_1 (week days vs. week end) due to insignificant correlation with SEVERITYCODE_1
- ST_COLCODE_1_1 due to low correlation

Pearson Correlation Heat Map on 2nd selection



Thus the features is leaner with only 7 features with to be used to build the Predictive Model.



	SEVERITYCODE_1
SEVERITYCODE_1	1.000000
SDOT_COLCODE_1	0.243995
JUNCTIONTYPE_1	0.221725
LIGHTCOND_1	0.167100
HITPARKEDCAR_1	0.122408
COLLISIONTYPE_1	0.111033
UNDERINFL_1	0.080007
WEATHER_1	0.027716

4.0 Predictive Model

The Model to predict factors that influence the increase in collision involving bodily injury must satisfy 3 requirements; measurability & ease in understanding and ability to produce actionable insight.
The outcome of the analysis is not only to produce a predictor of future collision but also to provide insight for City Government to do corrective action.

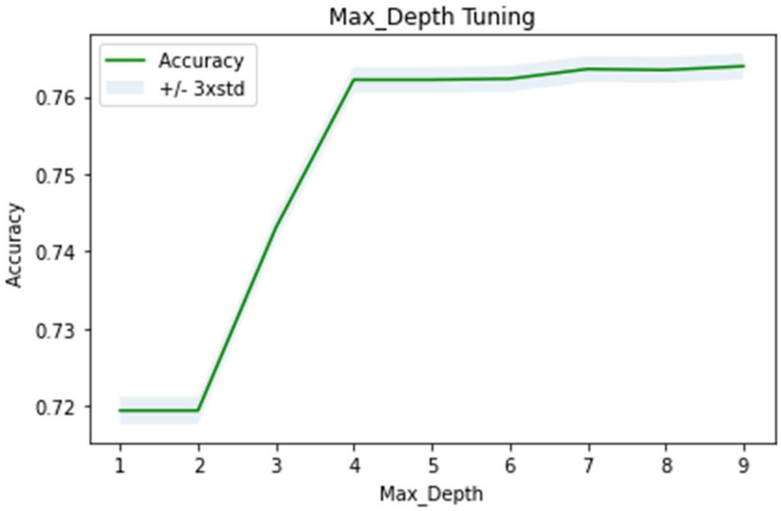
For the reasons above, the model used will be Decision Tree Model.

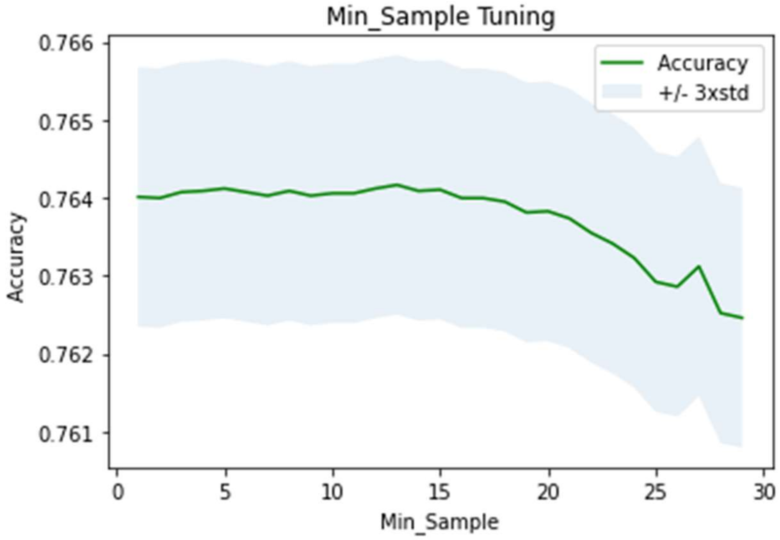
4.1 Model Training

Define Features

To Fit Decision Tree, the Features will be on-hot-encoded and sorted by the incident date to allow splitting between Data used for model training/testing and Data used to validate the result

As this is a Decision Tree Model with Entropy method, the model will be tuned on 2 parameters, Model Max Depth & Model Minimum Samples in leaf.

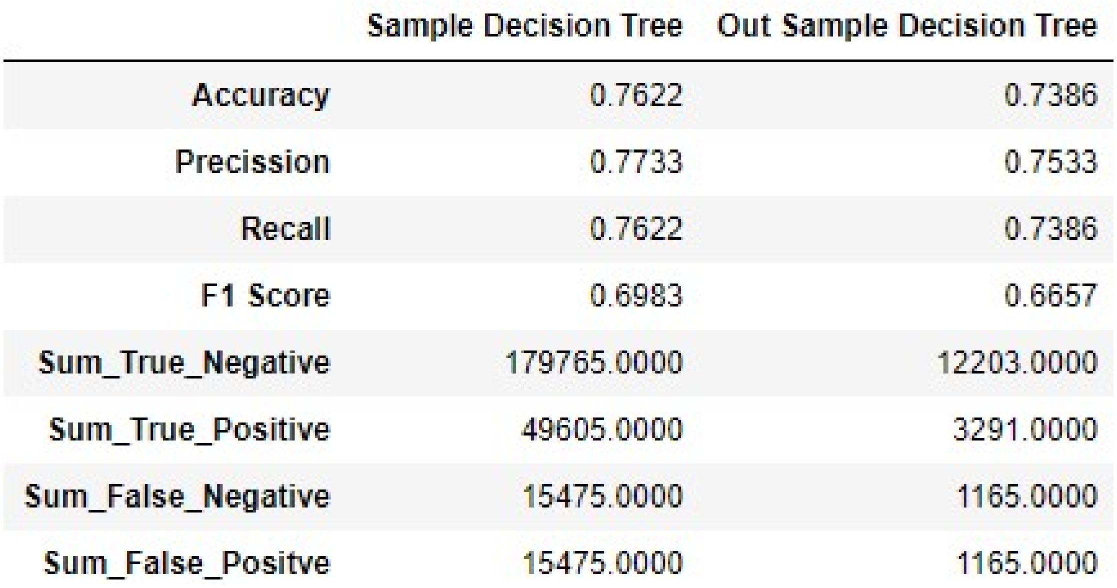




Max Depth Tuning using n Max Depth Iteration:
The Analysis indicates that after Max Depth of 4 there are no significant gain in accuracy. Thus Max Depth 4 is the most optimal Max Depth

Min_Sample Tuning using Maximum Iteration – n Iteration:
The Analysis indicates that after the 20th iteration, the Accuracy decrease significantly. Therefore the most optimal Minimum sample in a leaf is (30-20) = 10.

Based on the finding in the previous slide the Decision Tree will be configured using Max Depth of 4 and minimum sample in leafs of 10. In the same Process, the Model Performance is evaluated using Accuracy Score, Precision, Recall & F1 Score. The True & False prediction metrics also present to illustrate the Data Volume.



	Sample Decision Tree	Out Sample Decision Tree
Accuracy	0.7622	0.7386
Precission	0.7733	0.7533
Recall	0.7622	0.7386
F1 Score	0.6983	0.6657
Sum_True_Negative	179765.0000	12203.0000
Sum_True_Positive	49605.0000	3291.0000
Sum_False_Negative	15475.0000	1165.0000
Sum_False_Positve	15475.0000	1165.0000

The Tuned Decision Tree on Sample Data gives 0.76 accuracy which currently the most optimal fitting. With both Precision and Recall as well as F1 Score hover around 70%, the model is optimal in term of having minimum false prediction.

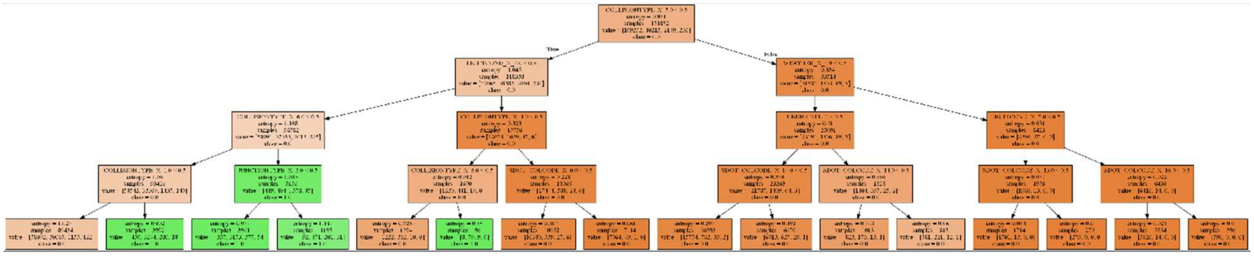
5.0 Decision Tree Model Conclusion

There are observable loss in accuracy using the target out sample data of 2020 collision. However in overall the model Accuracy, Precision, Recall and F1 Score are only reduced by roughly 2%, in overall the model can be generalized to new data set.

The Decision Tree Model that was built on Seattle GIS Collision Data for all Collision since 2004 is optimal to predict bodily injury in future collisions. With minimum Information on the more severe injury which involves not only serious Injury but also Fatality, the Model is only able to Predict Bodily injury in General. However,

Having healthy Recall & Precision score, It safe to assume that condition the model deem to be non-bodily related is also accurate thus allow decision maker or stake holder to focus on conditions that lead to bodily injury without the need to invest too much effort in other conditions.

The Decision Tree chart below is generated to better understand the condition that lead to bodily injury.



The Decision Tree split immediately between COLLISIONTYPE_X_5.0 (involves 'Parked Car'). Bodily Injuries only exist in leafs where they don't involve hitting parked car. In general there 3 conditions that Bodily Injury will occur:

- In any Known light condition (LIGHTCOND_X_7.0 <= 0.5), does not involved parked Car (COLLISIONTYPE_X_6.0 <= 0.5) and the vehicle hitting other vehicle or person from an angle (COLLISIONTYPE_X_1.0 > 0.5).
    - This a sound finding since there are known hazards involving location where vehicle, pedestrian or cyclist can intersect at an angle.
- In any Known light condition (LIGHTCOND_X_7.0 <= 0.5), does not involved parked Car (COLLISIONTYPE_X_6.0 <= 0.5) and the incident happened on 'Mid-Block (but intersection related)' (JUNCTIONTYPE_X_3.0 > 0.5).
    - This finding is intuitively sound these locations are most likely not equipped with enough road safety infrastructure that can help prevent collisions.
- The last condition is in and undefined light condition (LIGHTCOND_X_7.0 > 0.5) and involved parked Car (COLLISIONTYPE_X_6.0 > 0.5). This condition is quite small in term of number of incident (only 96 incidents since January 2020 to August 2020).
    - The last finding also intuitively sound since one can picture a condition where the driver need to back his/her car in a poor light condition and ended up hitting other vehicle or person.

Based on the finding City of Seattle may consider the following actions to reduce collision involving bodily injury.

Place road signs or warning as well as speed bumps on locations that may have vehicles, cyclist or pedestrian intersect with each other. Such locations may include mid-blocks but the model does suggest only limited to mid-blocks.

Improve street lighting or put regulation for safer parking rule to minimize injury involving parked car. However since the incident is not as often, this suggestion may takes later priority.

6.0 Future Development

Even though the Decision Tree Model is sufficient to describe conditions of the more general bodily Injury with relatively accurate result, it still lacks the ability to describe conditions that lead to Serious Injury and fatality which is the ultimate goal of this analysis. The lack of ability to describe the more detailed conditions is due to insufficient information specific to collision that have serious Injury and Fatality.

To Improve the Model ability to describe conditions that resulted in Serious Injury and Fatality, a more granular information specific to the location or the chronology of the incident have to be included into Data. Such information may include, what was the speed of the driver when approaching the incident location? Were both the driver/cyclist and the pedestrian were distracted? was there any specific building nearby such as schools, churches, shopping centers or location that may have a lot of traffics, and any other information that could better describe both how the accident happened and what the condition like in the incident location.