Development of a Predictive Model for Automobile Collision Severity

Coursera Applied Data Science Capstone Course Final Project

Evan Franke – September 1, 2020

Introduction

Knowledge and information, while omnipresent in today's society, are only beneficial if they can be communicated to others and converted to action. One promising area is in data associated with traffic accidents. While we are likely decades away from autonomous vehicles dominating our roads, drivers do have more and more information resources available to them while operating a vehicle that can enable them to make better informed decisions - decisions that can directly lead to increased safety.

The issue that will be addressed by this project is whether data collected on traffic incidents, when combined with meaningful attributes and characteristics, can be translated into predictive models that can be utilized in real time to warn drivers of likely dangerous situations in time to mitigate, or completely avoid, potentially dangerous traffic scenarios. These models, if proven even moderately accurate, could readily be integrated into existing end-user mobile device apps, as well as future integrated software in vehicle software operating systems.

Data

The data used for this effort will be the collision data published by SDOT, the Seattle Department of Transportation, which includes all collisions recorded from 2004 to May 2020. The data include 37 attributes for each collision, as well as the "severity level" of the collision itself - either an "injury collision" or a "property damage only collision." Attributes include location, intersection type, vehicle and occupant counts, date and time, and environmental conditions. The complete data set consists of 194,673 rows of individual collisions with their associated attributes, some of which are blank.

We will look for the subset of attributes that are reliable predictors of collisions by examining the relationship of each variable to collision severity. Any relationships that appear non-random will be used to create a predictive model via machine learning algorithms. From an early review of the available attributes, we expect fields such as "ADDRTYPE" (block / intersection / alley), "INCDTTM" (weekdays versus weekends, and commuting times), "ROADCOND" (wet / dry / ice and others), and "LIGHTCOND" (degrees of visibility) will likely be included in the model. There are many other descriptive fields available, but many of these would be difficult to translate into mathematical or categorical variables. The bulk of this effort is expected to be the discovery of which of the available attributes will ultimately be useful for our modeling.

Methodology

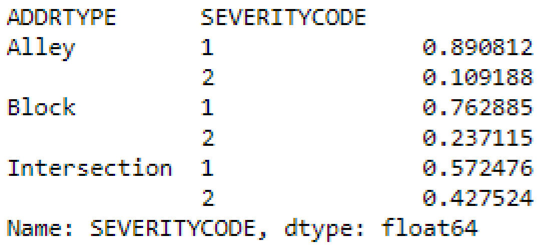
The “Data-Collisions.csv” file from Cognitive Class was downloaded into a data frame, which provided an opportunity to examine the data in detail. Severity Code is the first column, and consisted of only two different results: category 1 (indicating a collision with property damage only) and category 2 (an injury collision). Since these are the only two results recorded, it was decided that the purpose of this model should be to predict the occurrence of an injury collision based on ancillary attributes that existed at the time of the collision.

Omitted Fields

Database fields related to the attributes, and not descriptive of the event, were omitted from this effort. This includes OBJECTID, INCKEY, COLDETKEY, STATUS, INTKEY, SDOTCOLNUM, ST_COLCODE, SEGLANEKEY, and CROSSWALKKEY. Location data was also decided to be ignored, although there could possibly be a solid follow-up effort that perhaps directs the user to “safe” geographical alternatives in the event that an injury collision is predicted. Based on this scope limitation, X and Y coordinates, LOCATION, and JUNCTIONTYPE were ignored. Finally, any details of the collision itself (beyond severity) were ignored because the purpose of this tool will be to predict the event itself, and not the details or collateral damage thereof. For this reason, fields like COLLISIONTYPE, SDOT_COLDESC, ST_COLDESC and HITPARKEDCAR were disregarded.

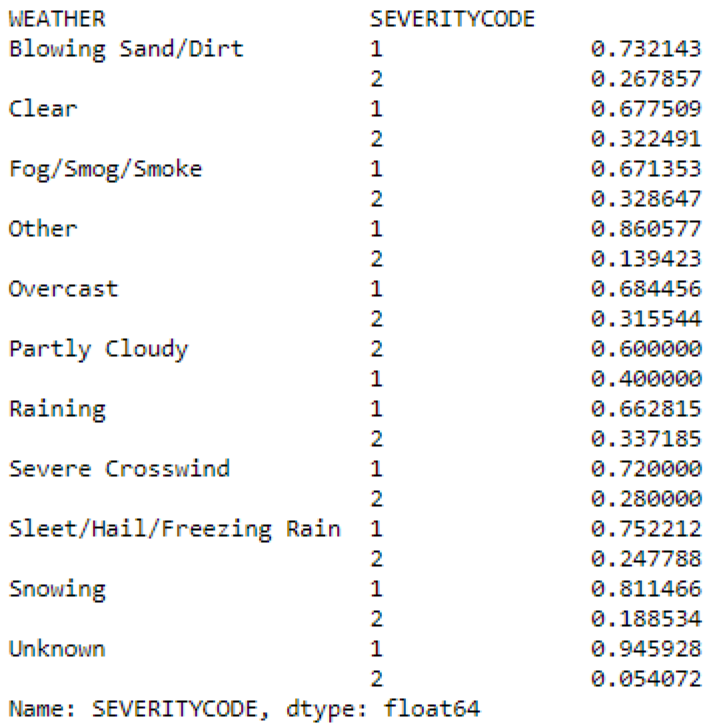
Selected Fields

ADDRTYPE – Alley, Block or Intersection. Although this is a descriptor of a location of sorts, the idea was to bring in this field as a potential indicator of an injury collision. The existence of this field could also be used for rudimentary injury severity mitigation instruction. The results were queried to ensure that a mix of results existed, shown below.



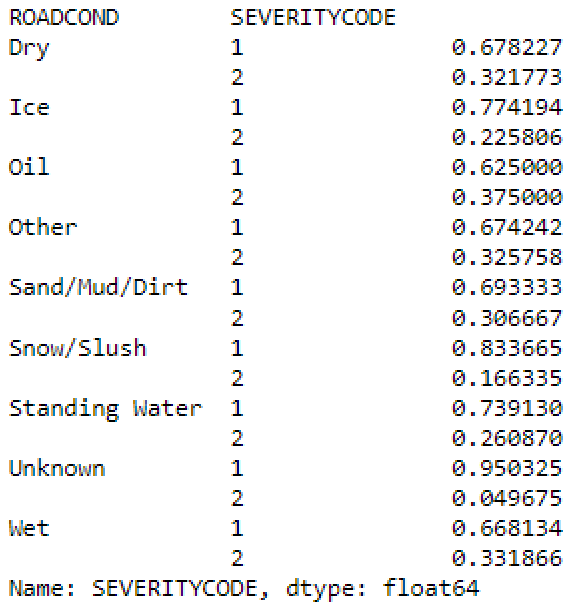
```
ADDRTYPE      SEVERITYCODE
Alley         1               0.890812
              2               0.109188
Block         1               0.762885
              2               0.237115
Intersection  1               0.572476
              2               0.427524
Name: SEVERITYCODE, dtype: float64
```

WEATHER – the weather reported at the time of the collision. This could easily be an input to the predictive tool from any mobile device. Again, the data was evaluated for a suitable mix of responses, with results shown below.



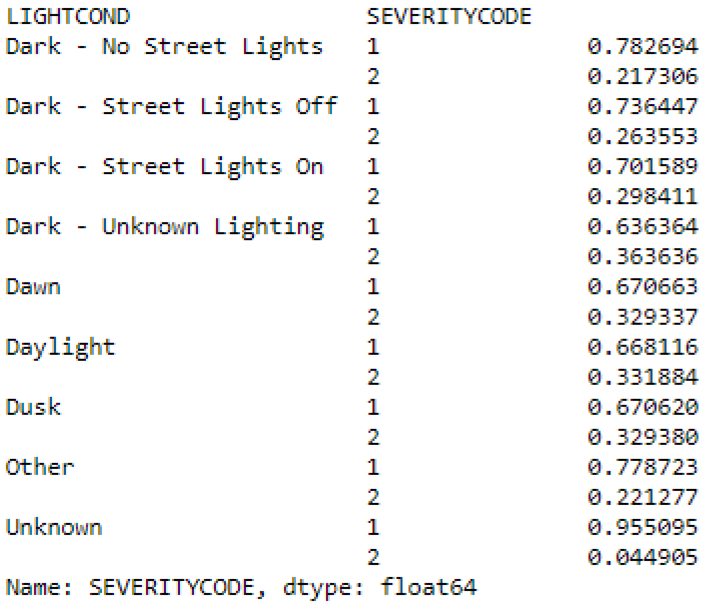
```
WEATHER                   SEVERITYCODE
Blowing Sand/Dirt         1               0.732143
                          2               0.267857
Clear                     1               0.677509
                          2               0.322491
Fog/Smog/Smoke            1               0.671353
                          2               0.328647
Other                     1               0.860577
                          2               0.139423
Overcast                  1               0.684456
                          2               0.315544
Partly Cloudy             2               0.600000
                          1               0.400000
Raining                   1               0.662815
                          2               0.337185
Severe Crosswind          1               0.720000
                          2               0.280000
Sleet/Hail/Freezing Rain  1               0.752212
                          2               0.247788
Snowing                   1               0.811466
                          2               0.188534
Unknown                   1               0.945928
                          2               0.054072
Name: SEVERITYCODE, dtype: float64
```

ROADCOND – the condition of the road at the time of the event, also likely available as a live input to the algorithm. Data evaluation is included below.



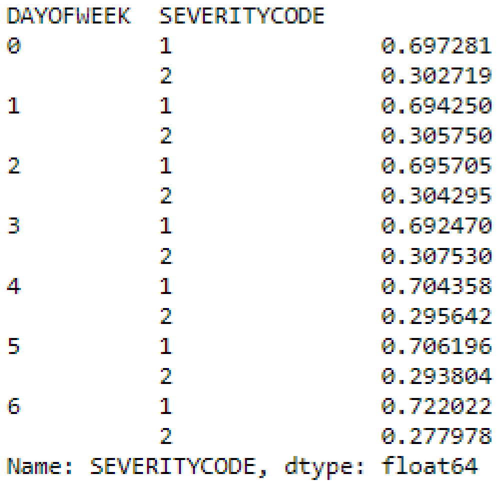
```
ROADCOND        SEVERITYCODE
Dry             1               0.678227
                2               0.321773
Ice             1               0.774194
                2               0.225806
Oil             1               0.625000
                2               0.375000
Other           1               0.674242
                2               0.325758
Sand/Mud/Dirt   1               0.693333
                2               0.306667
Snow/Slush      1               0.833665
                2               0.166335
Standing Water  1               0.739130
                2               0.260870
Unknown         1               0.950325
                2               0.049675
Wet             1               0.668134
                2               0.331866
Name: SEVERITYCODE, dtype: float64
```

LIGHTCOND – the environmental lighting at the time of the event, available as either an input from other mobile data sources, or as an estimate given the time of day combined with weather reports. Data evaluation follows.



```
LIGHTCOND                 SEVERITYCODE
Dark - No Street Lights   1               0.782694
                          2               0.217306
Dark - Street Lights Off  1               0.736447
                          2               0.263553
Dark - Street Lights On   1               0.701589
                          2               0.298411
Dark - Unknown Lighting   1               0.636364
                          2               0.363636
Dawn                      1               0.670663
                          2               0.329337
Daylight                  1               0.668116
                          2               0.331884
Dusk                      1               0.670620
                          2               0.329380
Other                     1               0.778723
                          2               0.221277
Unknown                   1               0.955095
                          2               0.044905
Name: SEVERITYCODE, dtype: float64
```

INCDTTM – the date and time of the collision. This was converted into "DAYOFWEEK" in order to determine if there was any relation to injury events such as the influence of weekday versus weekend traffic patterns, and traffic times related to commuting.



```
DAYOFWEEK  SEVERITYCODE
0          1               0.697281
           2               0.302719
1          1               0.694250
           2               0.305750
2          1               0.695705
           2               0.304295
3          1               0.692470
           2               0.307530
4          1               0.704358
           2               0.295642
5          1               0.706196
           2               0.293804
6          1               0.722022
           2               0.277978
Name: SEVERITYCODE, dtype: float64
```

Data Preparation

All the above fields seemed promising, so the next step was to extract just the fields of interest, and convert the categorical information into a format usable for machine learning by using one hot encoding algorithms. The result was a 33 column data structure consisting of numeric data. This was then transformed into scaled data so that no one data field would play an outsized role on the predicted result.

The data set, being quite large (>100.000 rows) was split into training and testing sets using a test size of 0.90 for reasons that will be described in the results section. This resulted in a training data set of nearly 20,000 collisions.

Selection of Machine Learning Tool

Four primary machine learning tools were expected to have the most success, considering that the input data was categorical in nature, and the result we desired was a code number. The four primary supervised Machine Learning tools were selected, specifically Support Vector Machine, K-Nearest Neighbor, Logistic Regression and Decision Tree. Each model would use the same training data set and generate predictions, and the resulting predictive accuracy will be used to select the best model.

Results

Data Training Size

Initially, a test size of 0.2 was used to split the data into training and testing subsets. This resulted in timeouts from the software, given the huge data size and the computations necessary to run the corresponding modeling. To test the timing, the test size was increased to 0.99, which resulted in very fast results (less than five seconds), with model accuracies around 58%. In order to achieve a balance between responsiveness and accuracy, a final test size of 0.90 was selected, which is still able to be executed in less than a minute, while achieving model accuracies around 70%.

Modeling Results

Using the four machine learning models discussed previously, test data (90% of the data set) was used to predict the occurrence of a level 2 severity traffic collision. All four methodologies were able to be used with the data provided, with the resulting accuracies listed in the table below.

Machine Learning Model	Test Set Accuracy
Support Vector Machine (SVM)	0.661
K-Nearest Neighbor (KNN)	0.676
Logistic Regression	0.701
Decision Tree	0.693

Discussion

The results of this short endeavor are very promising. A straightforward selection of intuitive data inputs that are readily available in a mobile device, with very little transformation, can be used to fairly accurately predict (70% accuracy) the occurrence of injury collisions for automobiles. The “Logistic Regression” machine learning algorithm appears to be the most accurate, and performed the computation in less than two seconds. Based on these results, it is reasonable to expect that incorporating such a model into existing driving aids could reduce the occurrence of injury accidents by:

1. Monitoring a subset of environmental factors, readily available from ancillary live data sources. This is likely to include weather, road and lighting conditions, as basic day and time information.

2. Supplying this data to the algorithm, which can predict with reasonable accuracy when a collision that includes an injury is likely to occur.
3. Providing an alarm to the driver, either live as the environmental attributes change and trip threshold alarms, or as the driver begins their trip as an alert to avoid driving altogether or to take mitigating action to reduce the likelihood of collision (such as alternate routes with better road conditions or fewer intersections).

Further development of these models is certainly warranted, with potentially added data fields for other readily available data. An example of a potentially valuable addition would be live traffic information, which could be used for live instructions on alternate safer routes to mitigate risk.

Conclusion

As mentioned in the introduction, information, though omnipresent in today's society, is only beneficial if it can be communicated to others and converted to action. This is demonstrated by the results of this project, which used records of nearly 200,000 collisions from Seattle to create a model that can predict the occurrence of injury accidents with 70% accuracy. In this case, employing machine learning algorithms derived from this data and using readily-available mobile device inputs has a very real possibility to reduce injuries and improve people’s lives. This should not only be of interest to mobile application developers and automobile software integrators as a valuable tool, but especially to all members of society as a tool to improve public health.