Accident severity prediction in Seattle

Manlai Amarsaikhan

September 1, 2020

1 Introduction

1.1 Background

If you commute daily, it is not uncommon to come across a car accident. You may not have seen it firsthand, but you definitely have experienced traffic jams that are caused by one. Car accidents are a daily occurrence on city roads and highways; and accidents cost a lot, financially, psychologically and as lost-time. People try to minimize these costs before it is even incurred by buying insurance. So in essence, because we want to avoid the costs associated with a chance event, we pay a definite amount to protect ourselves against it.

If we can predict the likelihood if an accident at any given moment, we can avoid all the costs and pain related to accidents. However, that is outside the scope of this project. If we accept accidents as a fact of life, the next best solution is to predict how bad an accident will be, given the current weather, time and road condition. If we can predict that, we can drive more carefully or even change our route if we can.

1.2 Problem

Clearly, it is preferable to avoid a car accident (or a collision) altogether, but in the event of a collision (which most likely will result in some property damage), presumably, it is preferable to avoid bodily injury as well. This project aims to predict the severity of collisions in Seattle. Specifically, given location, time, weather and road conditions in Seattle, the project aims to predict whether a collision will result in property damage or a bodily injury.

1.3 Audience

Firstly, Seattle residents may be interested in this study as the observations and findings in this study impact their daily life. Second, it may be interesting for Seattle Department of Transportation as the findings in this study may be useful to plan and design better roads and city planning. Lastly, even though this study used data specific to Seattle, its findings may be of interest to residents in other large cities, as the characteristics of another large city may be similar to that of Seattle.

2 Data

2.1 Data source

The main data set used in this report is provided by Seattle Police Department (SPD) and recorded by Traffic Records. The original data set has records of all types of collisions under SPD jurisdiction between 2004 and 2019. However, for the purposes of this report, a subset of this data set - namely, records of collisions which resulted in (i) property damage only, or (ii) bodily injury - was used. The data set was updated weekly by the Seattle Department of Transportation (SDOT) Traffic Management Division, Traffic Records Group (contact person: SDOT GIS Analyst, contact email: DOT_IT_GIS@seattle.gov).

2.2 Data cleaning

The data set contained 194,673 samples with 37 features and 1 label.

Using REPORTNO, a feature that collected the police reports filed on the recorded collisions, it was found that 3 collisions had two separate reports filed on them. Since the data except for the report number (REPORTNO) are exactly the same, these samples were dropped.

Next, an analysis of missing values revealed that three features (INATTENTIONIND, PEDROWNOTGRNT, SPEEDING) had 84.7% - 97.6% missing values. However, from the description of the features, it became clear that all three are dummy variables that have value Y if True and have no value if False. Therefore, the data set was corrected to reflect that.

Feature UNDERINFL depicted whether or not a driver involved was under the influence of drugs or alcohol. Clearly by its description, it is binary variable but it took four values - Y, N, 0, 1 - therefore, it was corrected to match its binary nature.

Features ROADCOND, LIGHTCOND, JUNCTIONTYPE and WEATHER depict the road, light conditions, junction type and weather during collisions, respectively. All take UNKNOWN as a value, which effectively means the corresponding condition information was missing. Therefore, all UNKNOWN values were corrected to NULL to reflect missing values.

After the initial data cleaning, the data set contained 194,670 samples with 37 features and 1 label.

2.3 Feature selection

Initial inspections showed that some features were duplicate. For example, features SEVERITYCODE.1 and SEVERITYDESC were a copy and a description, respectively, of the label SEVERITYCODE; thus completely mimicked the label; and features REPORTNO, OBJECTID, INCKEY, COLDETKEY, SDOTCOLNUM were

all related to assigning a unique key to a collision (by different organizations that keep track of this data set); thus irrelevant to this study. All of these features were dropped.

Features COLLISIONTYPE, ST_COLDESC, SDOT_COLDESC, SDOT_COLCODE and ST_COLCODE all described a collision (by two different organizations); thus these are highly correlated with each other. Feature HITPARKEDCAR described whether a collision involved hitting a parked car. This information was already included in COLLISIONTYPE. Therefore, COLLISIONTYPE was kept and the others were dropped.

Features X, Y, LOCATION, INTKEY, EXCEPTRSNDESC, EXCEPTRSNCODE all depicted location information. X and Y are numerical values that depicts the latitude and longitude of collisions, whereas LOCATION describes the general location of collisions. INTKEY depicts keys that corresponds to the intersection associated with a collision if a collision occurred at an intersection. EXCEPTRSNDESC and EXCEPTRSNCODE are dummy variables (one is a numerical value while the other is its description) that describe whether location information for a collision exists or not. X, Y were more useful and and the latter three became expendable; hence, these were dropped.

Feature INCDTTM depicts the date and time of an incident, while INCDATE depicts the date of an incident. Therefore, INCDTTM contains more information, thus, INCDATE was dropped.

Feature STATUS is an administrative feature that takes MATCHED and UNMATCHED as values. It is irrelevant to the study; thus was dropped.

Features CROSSWALKKEY and SEGLANEKEY depict keys for the crosswalk and the lane segment, respectively, in which a collision occurred. However, these contain 98% - 98.6% zero value; therefore do not contain important information. Both were dropped.

Table 1 summarizes the cleaning.

Lastly, all rows with missing values were dropped. After the feature selection, the data set contained 165,184 observations with 17 features which are summarized in Table 2:

Table 1: Initial feature selection

Kept features	**Dropped features**	**Reasons for dropping**
SEVERITYCODE	SEVERITYCODE.1, SEVERITYDESC	All three are equivalent.
	REPORTNO, OBJECTID, INCKEY, COLDETKEY, SDOTCOLNUM	All assign keys to collisions and closely correlate with each other.
COLLISIONTYPE	ST_COLDESC, SDOT_COLDESC, SDOT_COLCODE, ST_COLCODE, HITPARKEDCAR	First 4 describe a collision and closely correlate with each other. HITPARKEDCAR depicts whether a collision involved hitting a parked car, an information that is contained in COLLISIONTYPE.
X, Y ADDRTYPE	LOCATION, INTKEY, EXCEPTRSNDESC, EXCEPTRSNCODE	X,Y, INTKEY and LOCATION depict locations of collisions, but X, Y are more comprehensive since these are coordinates. The latter two are dummy variables that denote missing location information.
INCDTTM	INCDATE	Both depict dates of collisions but INCDTTM also includes time.
	STATUS, CROSSWALKKEY, SEGLANEKEY	Uninformative data

Table 2: Selected features

Features	**Description**
SEVERITYCODE	The code that corresponds to the severity of the collision: 2-injury, 1-prop damage
X, Y	The coordinates of a collision
ADDRTYPE	Collision address type: Alley, Block, or Intersection
COLLISIONTYPE	Collision type: A categorical variable taking 10 unique values
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.
INCDTTM	The date and time of the incident.
JUNCTIONTYPE	Category of junction at which collision took place: Takes 6 unique values
WEATHER	A description of the weather conditions during the time of the collision: Takes 9 unique values
ROADCOND	The condition of the road during the collision: Takes 8 unique values
LIGHTCOND	The light conditions during the collision: Takes 8 unique values
INATTENTIONIND	Whether or not collision was due to inattention.(Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.(Y/N)
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)

3 Methodology

3.1 Exploratory data analysis

3.1.1 Relationship between SEVERITYCODE and ADDRTYPE

Since both SEVERITYCODE and ADDRTYPE are categorical variables, it is best to analyze their relationship using a contingency table. Pearson's χ^2-test revealed that the variables are dependent and ADDRTYPE is a viable dependent variable ($\chi^2 = 7546.62$, degrees of freedom= 2, $p = 0.0$). Analysis of the expected frequencies showed that the observed number of collisions occurring at an intersection that results in an injury was significantly higher than expected.

Table 3 shows the contingency table for observed frequencies.

Table 3: Observed frequencies of SEVERITYCODE and ADDRTYPE

	ADDRTYPE		
SEVERITYCODE	**Alley**	**Block**	**Intersection**
1	669	96829	37251
2	82	30094	27819

Table 4 shows the expected frequencies:

Table 4: Expected frequencies of SEVERITYCODE and ADDRTYPE

	ADDRTYPE		
SEVERITYCODE	**Alley**	**Block**	**Intersection**
1	525.03	88732.97	45491
2	225.97	38190.03	19579

Based on the findings from this analysis, ADDRTYPE was updated into a dummy variable that takes value 1 if the collision occurred at an intersection and 0 otherwise.

3.1.2 Relationship between SEVERITYCODE and COLLISIONTYPE

Pearson's χ^2-test revealed that the variables are dependent and COLLISIONTYPE is a viable dependent variable ($\chi^2 = 41075.56$, degrees of freedom= 9, $p = 0.0$). Analysis of the expected frequencies showed that the observed number of collisions that results in an injury was significantly higher than expected if the collisions involved cycles, pedestrian, getting rear ended or sideswiped. On the other hand, collisions involving parked car had significantly fewer observations resulting in an injury.

Based on the findings from this analysis, COLLISIONTYPE was further updated into a dummy variable that takes value 0 if the collision type is parked car, right turn, sideswipe or other, and 1 otherwise.

Table 5: Observed frequencies of SEVERITYCODE and COLLISIONTYPE

SEVERITYCODE	parked car, right turn, sideswipe or other	Angles, cycles, head on, left turn, pedestrian, or rear ended
1	81365	55119
2	11889	46297

Table 6: Expected frequencies of SEVERITYCODE and COLLISIONTYPE

SEVERITYCODE	**parked car, right turn, sideswipe or other**	**Angles, cycles, head on, left turn, pedestrian, or rear ended**
1	65380.79	71103.21
2	27873.21	30312.79

3.1.3 Relationship between SEVERITYCODE and JUNCTIONTYPE

Pearson's χ^2-test revealed that the variables are dependent and JUNCTIONTYPE is a viable dependent variable ($\chi^2 = 8172.33$, degrees of freedom= 5, $p = 0.0$). Analysis of the expected frequencies showed that the observed number of collisions that results in an injury was significantly higher than expected if the collision occurred at an intersection and the collision was intersection related. On the other hand, collisions occurring mid-block without intersections had significantly fewer observations resulting in an injury. Since the insight to be obtained from this feature is already contained in ADDRTYPE, this feature was dropped when training the ML model.

The contingency tables are available on request.

3.1.4 Relationship between SEVERITYCODE and WEATHER

Pearson's χ^2-test revealed that the variables are dependent and WEATHER is a viable dependent variable ($\chi^2 = 94.0$, degrees of freedom= 9, $p = 0.0$). Analysis of the expected frequencies showed that the observed number of collisions that results in an injury was higher than expected if there was fog, smoke, smog, or rain. On the other hand, rather counter-intuitively, snow significantly negatively affected the likelihood of collisions resulting in an injury.

The contingency tables are available on request.

3.1.5 Relationship between SEVERITYCODE and ROADCOND

Pearson's χ^2-test revealed that the variables are dependent and ROADCOND is a viable dependent variable ($\chi^2 = 185.73$, degrees of freedom= 7, $p = 0.0$). Analysis of the expected frequencies showed that the observed number of collisions that results in an injury was higher than expected if the road is wet.

Table 7: Observed frequencies of SEVERITYCODE and ROADCOND

ROADCOND	**Dry**	**Ice**	**Oil**	**Other**	**Sand** Mud Dirt	**Snow Slush**	**Standing Water**	**Wet**
1	84446	936	40	89	52	837	85	31719
2	40063	273	24	43	23	167	30	15754

Table 8: Expected frequencies of SEVERITYCODE and ROADCOND

ROADCOND	**Dry**	**Ice**	**Oil**	**Other**	**Sand** Mud Dirt	**Snow Slush**	**Standing Water**	**Wet**
1	84301.62	818.58	43.33	89.37	50.78	679.78	77.86	32142.66
2	40207.38	390.42	20.67	42.63	24.22	324.22	37.14	15330.34

3.1.6 Relationship between SEVERITYCODE and LIGHTCOND

Pearson's χ^2-test revealed that the variables are dependent and LIGHTCOND is a viable dependent variable ($\chi^2 = 284.07$, degrees of freedom= 7, $p = 0.0$). Analysis of the expected frequencies showed that the observed number of collisions that results in an injury was higher than expected if it is daylight or dusk, which makes sense as there is far more traffic during the day rather than the night.

Table 9: Observed frequencies of SEVERITYCODE and LIGHTCOND

LIGHTCOND	**Dark No Street Lights**	**Dark Street Lights Off**	**Dark Street Lights On**	**Dark Unknown Lighting**	**Dawn**	**Daylight**	**Dusk**	**Other**
1	1203	883	34032	7	1678	77593	3958	183
2	334	316	14475	4	824	38542	1944	52

Table 10: Expected frequencies of SEVERITYCODE and LIGHTCOND

LIGHTCOND	**Dark No Street Lights**	**Dark Street Lights Off**	**Dark Street Lights On**	**Dark Unknown Lighting**	**Dawn**	**Daylight**	**Dusk**	**Other**
1	1043.75	814.22	32940.11	7.47	1699.06	78864.89	4007.93	159.58
2	493.25	384.78	15566.89	3.53	802.94	37270.11	1894.07	75.42

3.2 Classification strategy

First of all, we created dummy variables from the categorical variables. That expanded the feature set to 54 features.

The sample was first split into sample set and validation set (with 4:1 ratio). The sample set was further split into train set and test set (with 3:1 ratio). Since, there were plans to use SVM and KNN, a scaler was initialized

to normalize the train set, and it was also used to transform both the test set and the validation set.

Second, the data set was an unbalanced one (110,508 property-damage vs 54,676 injury samples). So before training any ML model, it was necessary to balance the samples to better understand the structure of the data set. To that effect, oversampling method SMOTE was used. After running the SMOTE method, there were 66,332 property-damage, and 66,332 injury samples in the train set.

Accuracy and ROC AUC scores were chosen as metrics because for this particular problem, it is most desirable to accurately predict the severity of an accident, i.e. we are agnostic about the tradeoff between false-negatives and false-positives, since both are equally undesirable. A false-positive result implies that our model had incorrectly predicted an injury when in fact it is just a property damage. Since the data set is unbalanced in favor of the property damage sample, a dummy classifier that always predicts the majority would perform the best if we aim to minimize false-positives. The costs of this false prediction is that drivers will have spent an unnecessary amount of time driving away from the location of that particular collision. A false-negative would imply that our model had incorrectly predicted a property damage when in fact it is a more serious accident that results in an injury.

4 Results

We trained and tuned K-Nearest Neighbor, Decision Tree, Logistic Regression, Random Forest, Gradient Boosting Classifier, as well as a dummy classifier that predicts the majority label. Due of the size of our data and time constraint, we decided against using Support Vector Machine. The accuracy and ROC AUC scores are collected in the Table 11.

First a dummy classifier was trained to see our benchmark accuracy. The accuracy of the dummy classifier was 0.67 and the ROC AUC score on the predicted probabilities of the positive values was 0.5.

Next, we trained a Logistic Regression. Using GridSearch, we found that $C = 1$ was the best parameter value. On the test set using cross validation, we found that its accuracy was 0.728.

Then, we trained Gradient Boosting Classifier. Using GridSearch, we found that learning_rate=0.1, max_depth=2, max_features=15 were the best parameter values. On the test set using cross validation, we found that its accuracy was 0.735.

Lastly, we trained Random Forest Classifier. Using GridSearch, we found that max_depth=4, max_features=13 were the best parameter values. On the test set using cross validation, we found that its accuracy was 0.732.

Table 11: Classification results

	Dummy	**LogisticRegression**	**GradientBoost**	**RandomForest**
accuracy	0.668	0.728	0.735	0.732
ROC AUC	0.5	0.753	0.766	0.764

5 Discussion & Conclusion

As it can be observed from Table 11, Gradient Boosting Classifier (GBC) performs the best among the chosen classifiers. All of them perform better than the dummy classifier; albeit not as well as one would have hoped for. It is due to the fact that

Interestingly, in terms of ROC AUC score, all classifiers perform much better than the dummy classifier. On this front, GBC performs the best as well.

Before we trained any machine learning algorithm, we inspected each feature individually with respect to the label as well as other features to "weed out" bad features. Specifically, we combined several values to define dummy variables for the categorical variables. For example, instead of creating dummy variables for each possible value of ADDRTYPE, we defined only one dummy variable (INTERSECTION) - whether a collision occurred at an intersection or not - and used it in training the ML models. We suspect that in doing so, we threw away some important features that could have been useful.

Moreover, we suspect that the strength of many features depend on other feature values. For example, we found that collisions (that result in injury) were highly correlated with hour of day (collisions occurring between 7-9am and 3-8pm were more likely to end up with an injury than any other hour), i.e. rush hour traffic was correlated with collisions that result in injuries. One would expect that this variable would interact with road condition, i.e. rush hour traffic + fog might result in collisions that result in injuries. Due to time constraint, we were not able to inspect it further but further research in that direction might yield better accuracy on ML models.