PREDICTING CAR ACCIDENT SEVERITY

Kim-Cuong Nguyen

Report Date: 09/15/2020

1. Introduction

1.1. Background

Car accidents and traffic jam have undeniably always been a headache problem to all car drivers. Car accidents happen for a lot of reasons. Those can stem from drivers' own mistakes as carelessness, alcohol, or speeding. Besides, there are also cases that the drivers get involved into collisions due to uncontrollable factors. For example, drivers are more prone to accidents when it is raining, and the road becomes wet and slippery. Or, even though the weather is clear, accidents are more likely to happen at several locations due to the characteristics of those places that make it more difficult for drivers to watch out for other drivers.

1.2. Objective

Due to the likelihood of car accidents to happen along the roads and the negative impacts they create, I was motivated to build a model to predict the possibility of a car accident and its severity. Specifically, I will use the data science powers to build a predictive model that can warn drivers about the possibility of getting into a car accident and the potential severity, so that the drivers can drive more carefully or reroute if possible.

2. Data

2.1. Data Overview

To meet the objective, I employed a dataset that lists collisions from 2004 to May 2020 in Seattle, recorded by Seattle Department of Transportation (SDOT). This is comprised of all types of collisions that display at the intersection or mid-block of a segment. The dataset is updated weekly.

The dataset is comprised 194,673 observations that are details on collisions during the timeframe. 38 variables are included, of which 18 are categorical variables with lots of them being high cardinality.

Appendix provides a brief explanation on the variables.

2.2. Data Cleaning

2.2.1. Irrelevant/Redundant Features

By examining the meaning of each variable within the dataset, I decided to remove the following ones and keep the meaningful variables only.

- **SEVERITYCODE:** A code that corresponds to the severity of the collision. This column is the same as column 'SEVERITYDESC' which gives the detailed description of the severity of the collision. Therefore, either of the two columns can be removed.
- **OBJECTID, INCKEY, COLDETKEY, REPORTNO, INTKEY, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY:** These are identities assigned to the collisions by SDOT and have no meaning in predicting the accident; thus, can be removed.
- **STATUS**: There is no description available for this attribute. It has two values "Matched", and "Unmatched".
- **ADDRTYPE, LOCATION**: Detailed address of collision locations. This corresponds with the two columns of longitude and latitude. The columns of longitude and latitude are better inputs into the model because the location attribute is of high cardinality.
- **EXCEPTRSNCODE, EXCEPTRSNDESC:** The two columns give some notes about data as "not enough information", etc.
- **COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, INATTENTIONIND, UNDERINFL, PEDROWNOTGRNT, SPEEDING, HITPARKEDCAR**: The attributes give details on collision outcomes, while the models aim at predicting the accident. Thus, the variables will be removed.
- **INCDATE**: The date of accident. This is part of another variable "**INCDTTM**" that shows date and time of the accident.
- **JUNCTIONTYPE**: This attribute is similar to "**ADDRTYPE**".

The sample observations of the dataset are as below:

X	Y	SEVERITY-DESC	INCDTTM	WEATHER	ROAD-COND	LIGHT-COND
-122.323	47.70314	Injury Collision	3/27/2013 14:54	Overcast	Wet	Daylight
-122.347	47.64717	Property Damage Only Collision	12/20/2006 18:55	Raining	Wet	Dark - Street Lights On

Table 1. Sample data

2.2.2. Missing Values

The number of observations with missing values is insignificant per variables. The total number of observations with missing values is 10,506, accounting for about 5.4%. Therefore, I decided to remove all missing values from the dataset.

X	Y	SEVERITY-DESC	INCDTTM	WEATHER	ROAD-COND	LIGHT-COND
2.74%	2.74%	0%	0%	2.61%	2.57%	2.66%

Table 2. Missing values percentages

By further checking the values of three variables, WEATHER, ROADCOND, and LIGHTCOND, I found that the variables have the value 'Unknown' that is similar to missing values. The number of observations with this value is 17,462, accounting for 9.5% of the remaining observations, which is not significant. Therefore, I removed all these observations.

Because the original dataset is too large, leading to my failure to complete modelling due to a lack of computational capabilities and infrastructure resource, I decided to reduce the data size by using data from 2018 to present only, including 18,708 observations.

2.2.3. Data Transformation

Variable INCDTTM gives information on date and time of the collision. Instead of using this variable, I extracted this into three other variables to input into the models, including month of collision, day of week, and hour.

I converted categorical variables (WEATHER, ROADCOND, and LIGHTCOND) into dummy variables, and removed the values 'Other'. After the conversion, I checked correlation among all the variables and found nine pairs of variables with high correlation (absolute values over 0.5) and removed six variables. The variables below are highly correlated, and those in bold are removed from the dataset.

- WEATHER_Clear vs **WEATHER_Overcast**
- WEATHER_Clear vs **WEATHER_Raining**
- WEATHER_Clear vs **ROADCOND_Dry**
- WEATHER_Clear vs **ROADCOND_Wet**
- **WEATHER_Raining** vs **ROADCOND_Dry**
- **WEATHER_Raining** vs **ROADCOND_Wet**
- WEATHER_Snowing vs **ROADCOND_Snow/Slush**

- **ROADCOND_Dry** vs **ROADCOND_Wet**
- LIGHTCOND_Daylight vs **LIGHTCOND_Dark - Street Lights On**

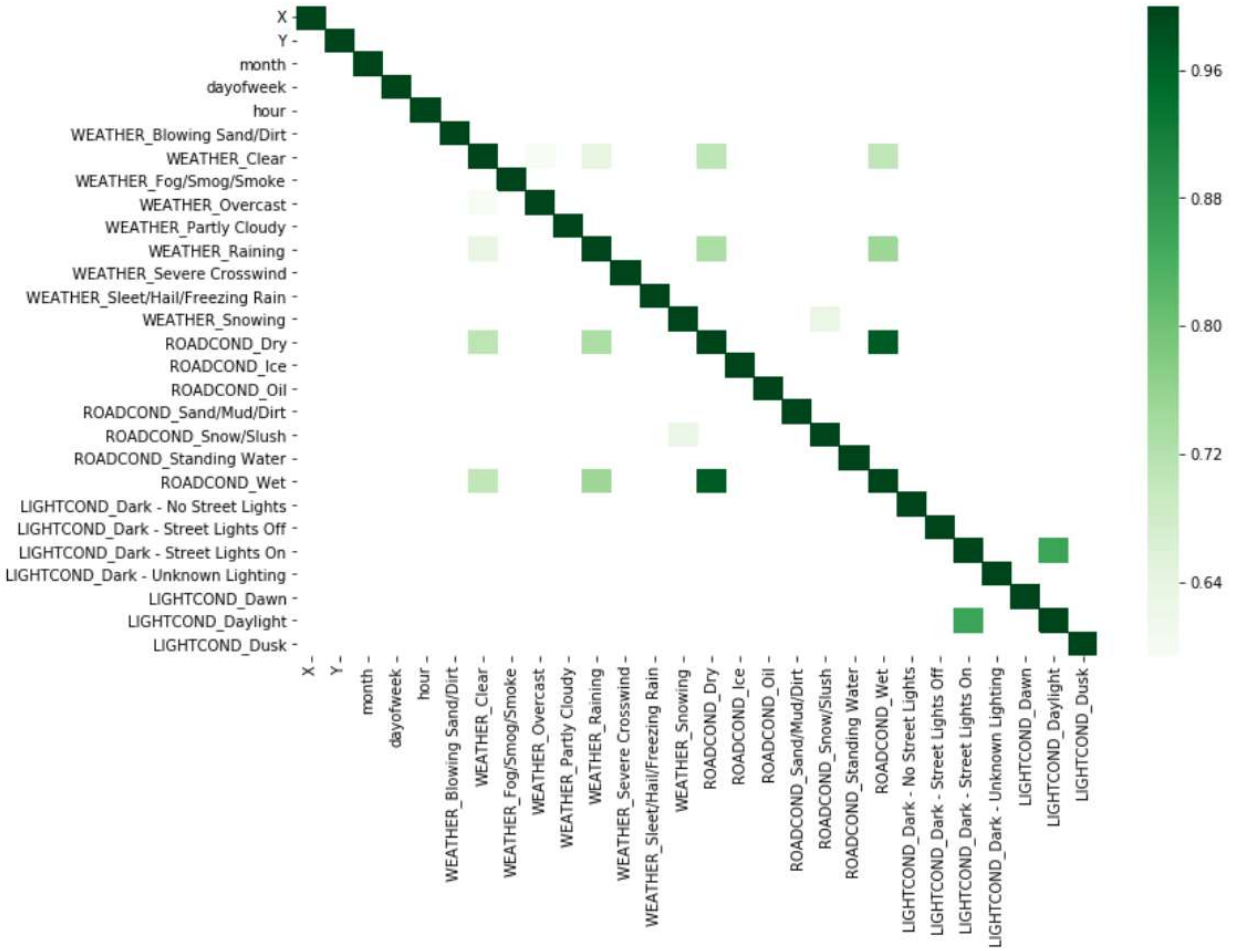

Figure 1. Heatmap of absolute correlation values above 0.5 among variables

The dataset contains numeric data on different scales. Therefore, I standardized the data before proceeding with modelling step.

2.3. Data Exploratory

The target variable of the modelling is SEVERITYDESC. There are two classes in the target variable: accidents with injuries, and those with property damage only. Some initial analysis was performed to find out relationship between independent variables and the target variable.

2.3.1. Car accidents vs day of week

First, it is noted that about two thirds of collisions are involved with property damage only while injuries incur in the remaining.

Accidents are more likely to happen during weekdays than weekends. This is understandable because weekdays are time when people have to travel more frequently to work and to school. Fridays are the time with the highest number of accidents during a week.

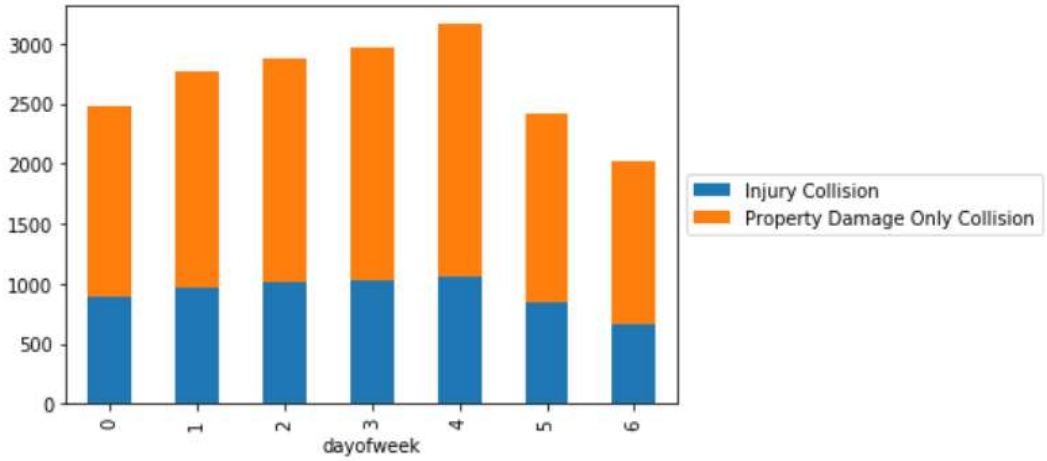

Figure 2. Collision frequency during the week

2.3.2. Car accidents vs time of a day

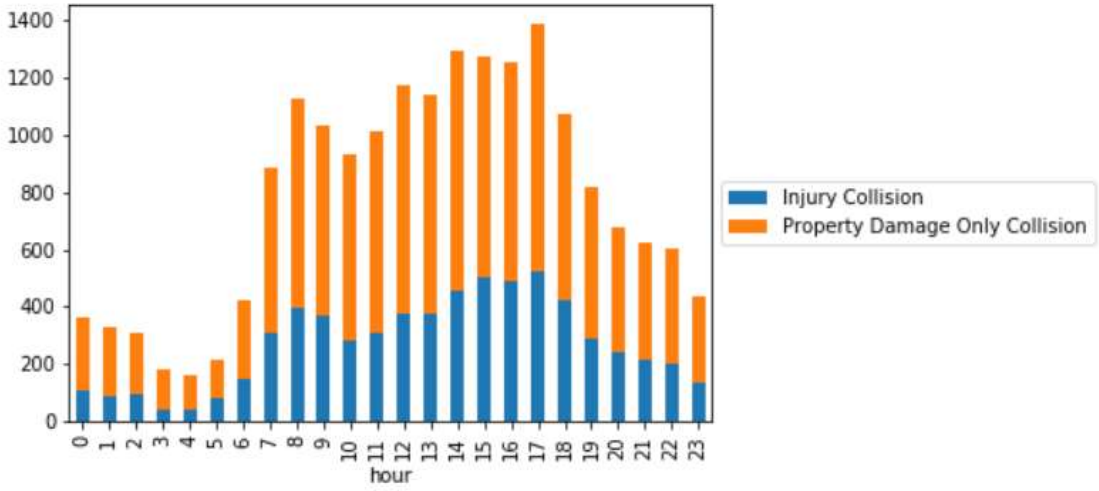

Figure 3. Collision frequency during a day

As expected, collisions are more likely to happen between 7am to 7pm – the time when people need to travel to work and school. Peak hours in the afternoon is the time when accidents are most likely to happen.

2.3.3. Car accidents vs month

There is no clear relationship between car accidents and months, though January is the month with the highest number of collisions during a year.

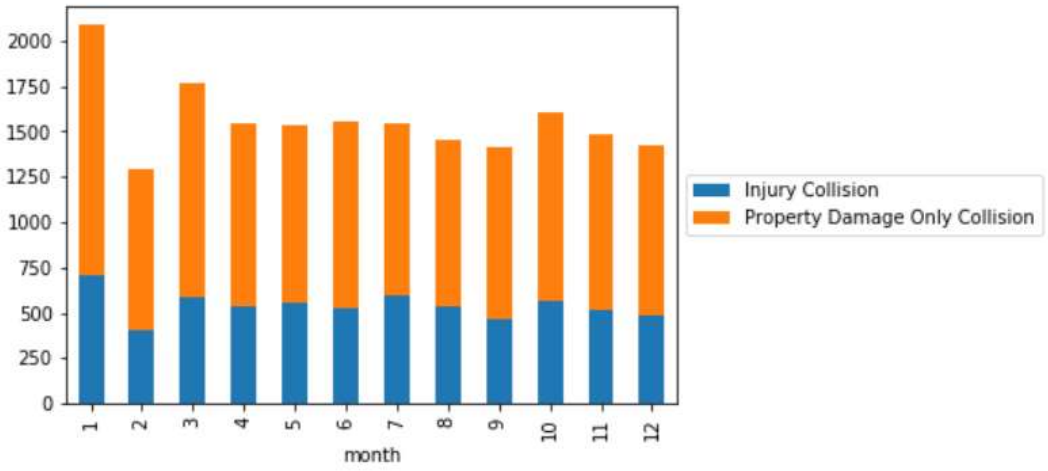

Figure 4. Collision frequency during a year

2.3.4. Car accidents vs weather, road, and light conditions

Though most reported accidents occur during normal weather, road, and light conditions, it is impossible to conclude that accidents cannot be attributed to these factors.

Most collisions occur when it is clear, overcast, or raining while few accidents happen with other weather conditions. The reason may be that people tend to find a shelter instead of driving when the weather is too bad, snowing, freezing rain, froggy, for example, thus, not many accidents were reported with these weather conditions.

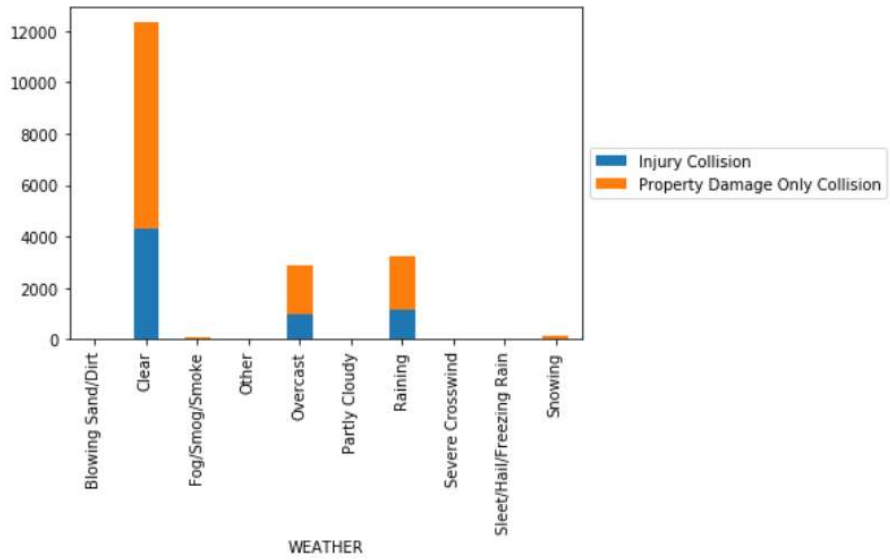

Figure 5. Collision frequency vs weather conditions

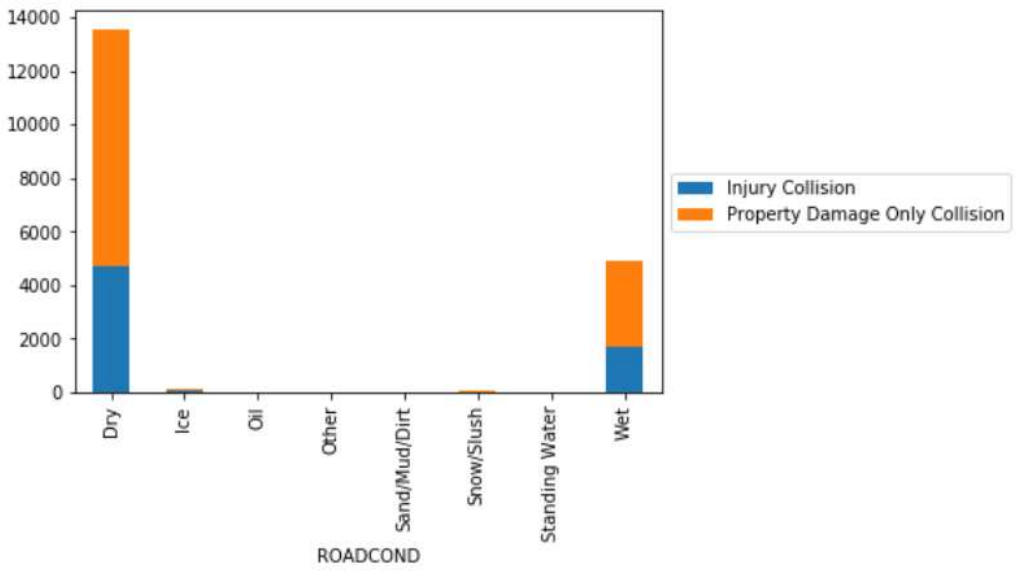

Figure 6. Collision frequency vs road conditions

Similarly, that few accidents were reported to be associated with icy, oily, snowy road does not imply that accidents do not happen when the road conditions are that bad. There are chances that drivers do not continue the trip on such roads.

This also holds true to light conditions. Accidents can happen anytime during the day or at night with street lights on.

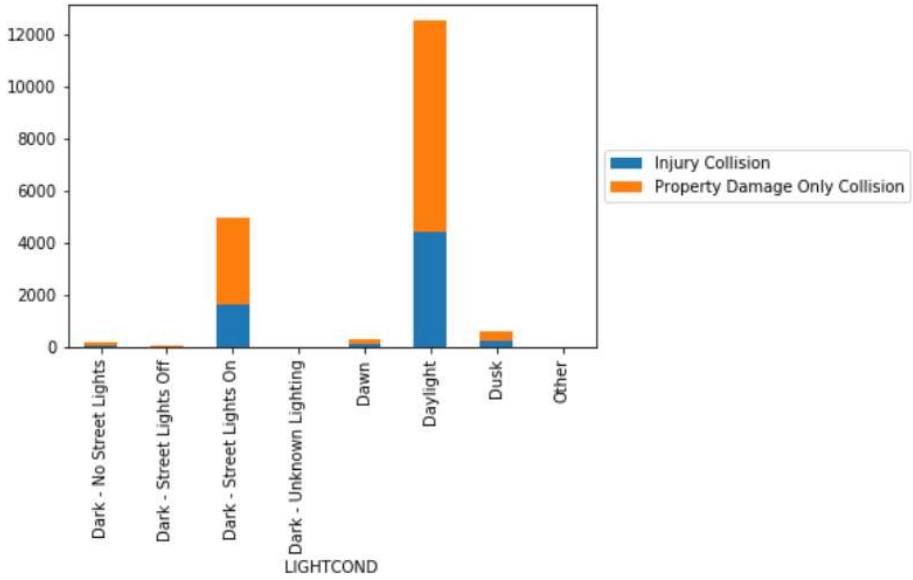

Figure 7. Collision frequency vs light conditions

3. Modelling

I tried different techniques to find the best model. I chose to split the dataset into training data and testing data by 75%-25%. I chose 75%% for training set because I believe the dataset with 18,708 observations is big enough and 75% of the dataset is good enough to train the model.

The data is imbalanced with one class being about half of the other one. However, the imbalance is not so serious, so I did not oversample the data to fix this.

3.1. kNN

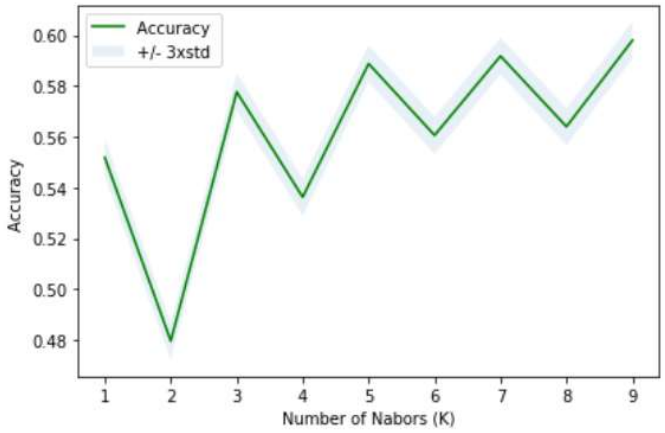

Figure 8. Model accuracy against k

Based on the plot, k = 9 gives the model with the best accuracy. By training the kNN model with k = 9, I got a model with accuracy levels for train and test set being 69.1% and 59.8% respectively. The model has an overfitting issue. Below is the confusion matrix on test data.

		Predicted Class	
		Injuries	Property
True Class	Injuries	289	1349
	Property	531	2508

3.2. Decision Trees

By training the model with different depth, I chose the depth being 9 to get the best model possible. The decision tree model gives an accuracy level for train and test data of 67.1% and 64.1% respectively.

Below is the confusion matrix on test set.

		Predicted Class	
		Injuries	Property
True Class	Injuries	103	1535
	Property	146	2893

3.3. Support Vector Machine

By trying different kernel, I chose the model using sigmoid kernel. Even though this model did not produce the best accuracy levels, it did better in classifying between the two predicted classes. Other models give higher accuracy but failed to predict accidents with injuries. Meanwhile, this class is of great importance. Between the two classes, a false prediction of accidents with injuries is still more important than a true prediction of accidents with property damage only.

The model trained by support vector machine with sigmoid kernel gives accuracy levels of train and test set as 61% and 61.4% respectively. Confusion matrix on test set is as below:

		Predicted Class	
		Injuries	Property
True Class	Injuries	169	1469
	Property	335	2704

3.4. Logistic Regression

The model using logistic regression gives an accuracy level for train and test set of 65.5% and 64.9% respectively. The model could not predict accidents with injuries. Confusion matrix on test set is as below:

		Predicted Class	
		Injuries	Property
True Class	Injuries	0	1638
	Property	3	3036

4. Conclusion

The table 3 summarizes performance of the selected classification models. The one with best performance per criterion is labelled in red.

Even though kNN and Logistic Regression each takes a lead in half of the criteria, the models in general are not good choices. Logistic regression model failed to predict

accidents with injuries. Therefore, despite this model could predict with a high overall accuracy, this is not the model I am looking for. Besides, kNN has a flaw that the model is a bit overfitted with low accuracy level on test set.

Comparing decision trees model and SVM model, I chose SVM model despite its lower accuracy. The reason is that this model is better at predicting accidents with injuries, which is of great importance in this situation.

Description		**kNN**	**Decision Trees**	**SVM**	**Logistic Regression**
Accuracy	Train	69.1%	67.1%	61%	65.5%
	Test	59.8%	64.1%	61.4%	64.9%
True	*Predicted*				
Injuries	Injuries	289	103	169	0
Injuries	Property	1349	1535	1469	1638
Property	Injuries	531	146	335	3
Property	Property	2508	2893	2704	3036

Table 3. Performance of classification models. Best performance labeled in red.

5. Recommendation

I think predicting possibility of collision and its severity is of great importance to warn drivers to be more careful and avoid bad situation. Because this model takes inputs as longitude and latitude of locations where accidents often occur, weather, road conditions, and time, one suggestion is that it can become a built-in function of GPS devices. Drivers can learn about places with high likelihood of accidents on their planned route and find ways to reroute if possible.

6. Limitation

This modelling and analysis still have several limitations as below.

- The accuracy level is still low of over 61%. One of possible reason is that the amount of data used in the modelling section is still small because I could not use the full dataset due to limitations in terms of computational capabilities and infrastructure resources. Plus, the inputs into the models are mostly external variables like weather, road condition, locations, etc. while many accidents happen due to drivers' carelessness or other drivers' lack of attendance. A lack of variables to explain accident possibility can be a reason for model's low accuracy.
- The dataset provides details of collisions in Seattle only while the model takes into account of location information; thus, this model cannot be used to predict accident

possibility and severity in locations outside of Seattle.

- The target variables have only two classes – accidents with injuries and with property damage only, thus, limiting the capability of the model to predict other scenarios.

APPENDIX – Variable Definitions

X	Longitude of the collision location
Y	Latitude of the collision location
OBJECTID	ESRI unique identifier
INCKEY	A unique key for the incident
COLDETKEY	Secondary key for the incident
REPORTNO	Another key for the incident
STATUS	N/A
ADDRTYPE	Collision address type: Alley, Block, Intersection
INTKEY	Key that corresponds to the intersection associated with a collision
LOCATION	Description of the general location of the collision
EXCEPTRSNCODE	Code to indicate data status
EXCEPTRSNDESC	Description of the data status
SEVERITYCODE	A code that corresponds to the severity of the collision
SEVERITYDESC	A detailed description of the severity of the collision
COLLISIONTYPE	Collision type
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	The number of vehicles involved in the collision. This is entered by the state.
INCDATE	The date of the incident.
INCDTTM	The date and time of the incident.
JUNCTIONTYPE	Category of junction at which collision took place.
SDOT_COLCODE	A code given to the collision by SDOT
SDOT_COLDESC	A description of the collision corresponding to the collision code.
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.
WEATHER	A description of the weather conditions during the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	A number given to the collision by SDOT.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	A code provided by the state that describes the collision.
ST_COLDESC	A description that corresponds to the state's coding designation. This is the target variable.

SEGLANEKEY	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)