1. Introduction

1.1 Background

This project is to help the department of transport and traffic improve the quality of specific warnings given to drivers. The traffic accidents can lead to great social inconvenience and loss, so the departments of traffic around the world are looking for effective measures to prevent the occurrences. One of the measures is to give warnings to drivers of the severity of car accidents potentially involved in.

1.2 Problem

To make the drivers more aware of the potential danger in their surroundings, various road and traffic departments are looking for critical factors affecting the severity of car collisions. By identifying these factors, they can take more effective preventive measures before car accidents happen.

The project will build classification models and help predict the severity level of car accidents based on weather, the road conditions and other relevant aspects.

1.3 Interests

On the one hand, traffic and road departments can put warning signs and adjust them when necessary. On the other hand, it may be integrated to the mobile devices which can inform drivers of the potential danger of collisions in real time.

2. Data acquisition and cleaning

2.1 Data sources

The collision dataset is collected by SDOT Traffic Management Division, Traffic Records Group in Seattle, which includes all collisions provided by SPD and recorded by Traffic Records from 2004/01/01 to 2020/05/20.

2.2 Preliminary Feature selection

There are 6 variables selected to be explored and potentially used as input in modelling. The reason for choosing these features is that they can be identifies before collisions happen.

No.	Variable	Description	Type
1	ADDRTYPE	Collision address type	Categorical
2	JUNCTIONTYPE	Category of junction	Categorical
3	WEATHER	A description of the weather conditions	Categorical
4	ROADCOND	The condition of the road	Categorical
5	LIGHTCOND	The light conditions	Categorical

6	SPEEDING	Whether speeding involved	Categorical

3. Exploratory Data Analysis

Data understanding: There are 194673 records of collisions and 38 attributes in the dataset.
The target variable is SEVERITYCODE, which is unbalanced as 136485 records belong to category 1 (property damage) and 58188 records belong to category 2 (injury).

3.1 ADDRTYPE

[Value count]

Block	126926
Intersection	65070
Alley	751

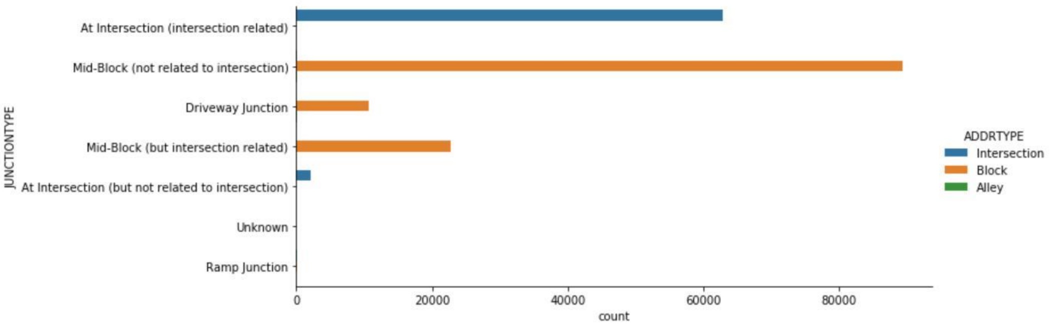
[Encode]
Block = 0
Intersection = 1
Alley = 2

3.2 JUNCTIONTYPE

[Value count]

Mid-Block (not related to intersection)	89800
At Intersection (intersection related)	62810
Mid-Block (but intersection related)	22790
Driveway Junction	10671
At Intersection (but not related to intersection)	2098
Ramp Junction	166
Unknown	9

We find it is a **subset of ADDRTYPE**, so we **drop** it



3.3 WEATHER

We find there are some similarities between several values and can be further integrated.
[Encode]
Clear = 0

Overcast = 1; Partly Cloudy = 1
Fog/Smog/Smoke = 2; Blowing Sand/Dirt = 2; Severe Crosswind = 2
Raining = 3; Sleet/Hail/Freezing Rain = 3; Snowing = 3
Unknown = 999; Other = 999

3.4 ROADCOND

We find there are some similarities between several values and can be further integrated.
[Encode]
Dry = 0
Sand/Mud/Dirt = 1; Standing Water = 1; Wet = 1
Oil = 2; Ice = 2; Snow/Slush = 2;
Unknown = 999; Other = 999

3.5 LIGHTCOND

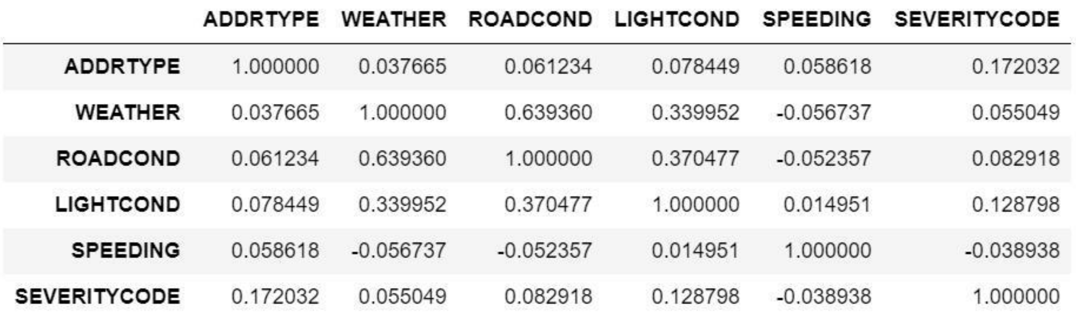
We find there are some similarities between several values and can be further integrated.
[Encode]
Daylight = 0
Dusk = 1; Dawn = 1; Dark - Street Lights On = 1
Dark - No Street Lights = 2; Dark - Street Lights Off = 2
Unknown = 999; Other = 999, Dark - Unknown Lighting = 999

3.6 SPEEDING

We assume nan value in SPEEDING means No speeding problems involved.
[Encode]
N = 0
Y = 1

3.7 Correlation between variables

We further **drop** the two variables with smallest correlations with the target variable, **WEATHER** (0.055) and **SPEEDING** (-0.0389) to reduce the noise and simplify the model.



	ADDRTYPE	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	SEVERITYCODE
ADDRTYPE	1.000000	0.037665	0.061234	0.078449	0.058618	0.172032
WEATHER	0.037665	1.000000	0.639360	0.339952	-0.056737	0.055049
ROADCOND	0.061234	0.639360	1.000000	0.370477	-0.052357	0.082918
LIGHTCOND	0.078449	0.339952	0.370477	1.000000	0.014951	0.128798
SPEEDING	0.058618	-0.056737	-0.052357	0.014951	1.000000	-0.038938
SEVERITYCODE	0.172032	0.055049	0.082918	0.128798	-0.038938	1.000000

3.8 Feature Selection

We select three features after the data analysis and exploration.

Variable	Description	Type
ADDRTYPE	Collision address type	Categorical
ROADCOND	The condition of the road	Categorical
LIGHTCOND	The light conditions	Categorical

4. Predictive Modeling

We use various classification models to predict categorical target variable with only two values. Also, 3:7 test-train data is used to improve the prediction performances of models. Instead of using accuracy in the imbalanced dataset, we use f1 score to find the optimal model. (Only 30000 random samples selected to build the model due to system restrictions in this project, easy to crash)

4.1 K-Nearest Neighbor(KNN)

When k = 4, the model performs best.

4.2 Decision Tree

When depth = 1, the model performs best.

4.3 SVM

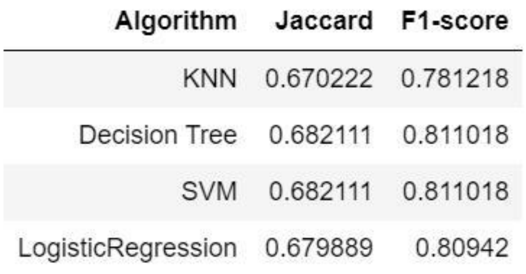
We try different kernels and find their performance is the same.

4.4 Logistic Regression

We try different solvers and find RBF (Radial Basis Function) and linear has the best performance.

4.5 Performance

The performances of different models are evaluated based on Jaccard similarity and F1-score. The higher they are, the performance of the model is better.
However, the depth of decision tree is only 1 but with the best performance. The model simply adheres to predict the level 1 severity and sacrifice the accuracy in predicting level 2 severity involving injury.



Algorithm	Jaccard	F1-score
KNN	0.670222	0.781218
Decision Tree	0.682111	0.811018
SVM	0.682111	0.811018
LogisticRegression	0.679889	0.80942

5. Conclusions

It seems the features in the imbalanced dataset are inadequate to distinguish car collisions involving injury from these only having property damage in the prediction. As more features to be developed and collected, it may be possible to forecast serious collisions with injury more accurately in the future.