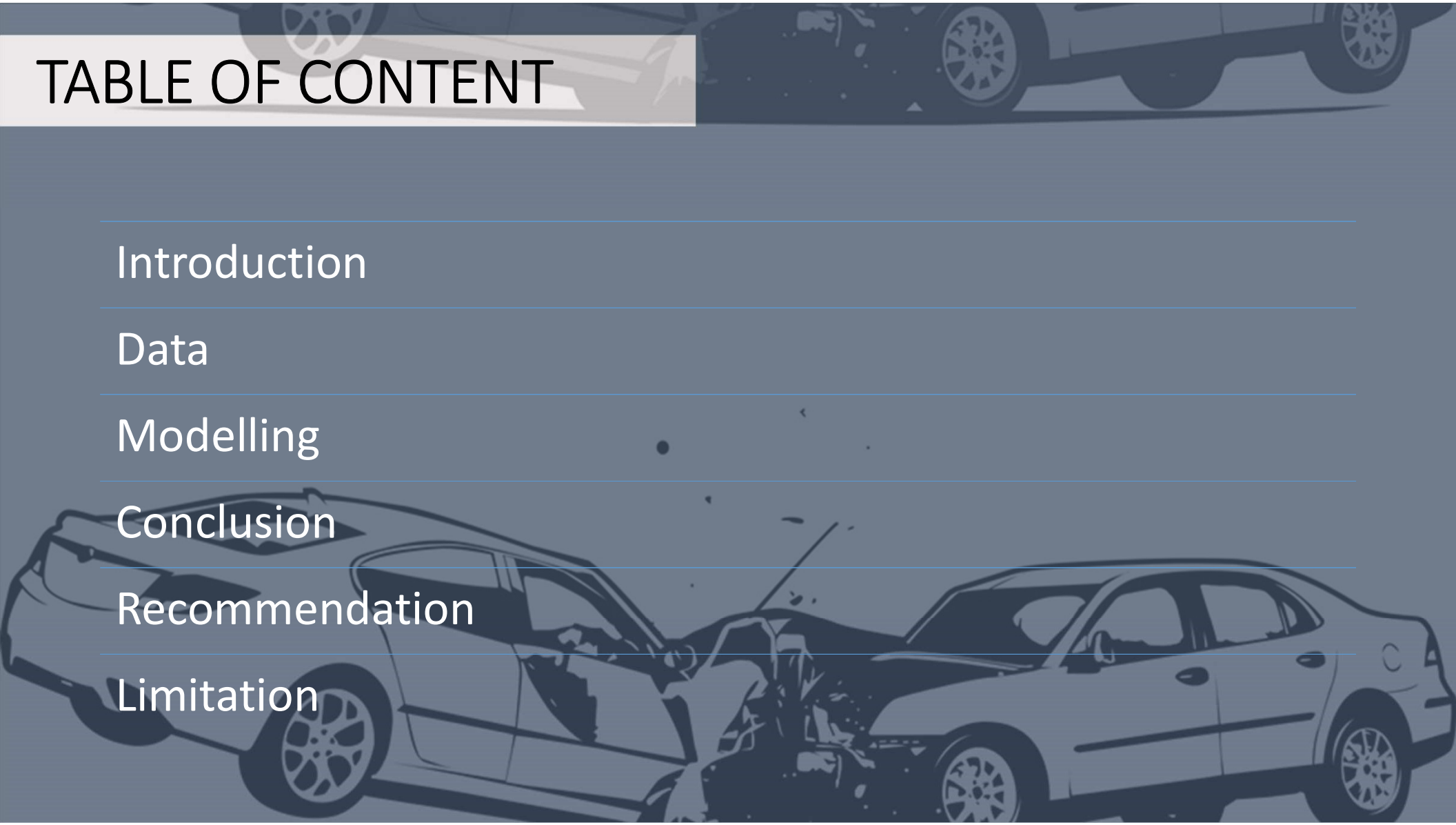
CAR ACCIDENT SEVERITY PREDICTION

Capstone Project

Kim-Cuong Nguyen

Report Date: September 15, 2020

TABLE OF CONTENT



INTRODUCTION

Background

- Car accidents and traffic jam - a headache problem to all car drivers.
- Car accidents happen for a lot of reasons.
- Car accidents bring about lots of negative consequences.

Objective

- Build a model to predict the possibility of a car accident and its severity.
- 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.

DATA SUMMARY

- Dataset of collisions from 2004 to May 2020 in Seattle, recorded by Seattle Department of Transportation (SDOT)
- 194,673 observations - 38 variables
- Remove 31 variables of three types: duplicates, collision identities assigned by SDOT, collision consequences
- Remove missing values and observations with values 'Unknown') – 15% of the total dataset
- Final variables: longitude and latitude of collision, target variable, date time of collision, weather, road, and light conditions when the collision occurs

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

Sample data

DATA PRE-PROCESSING

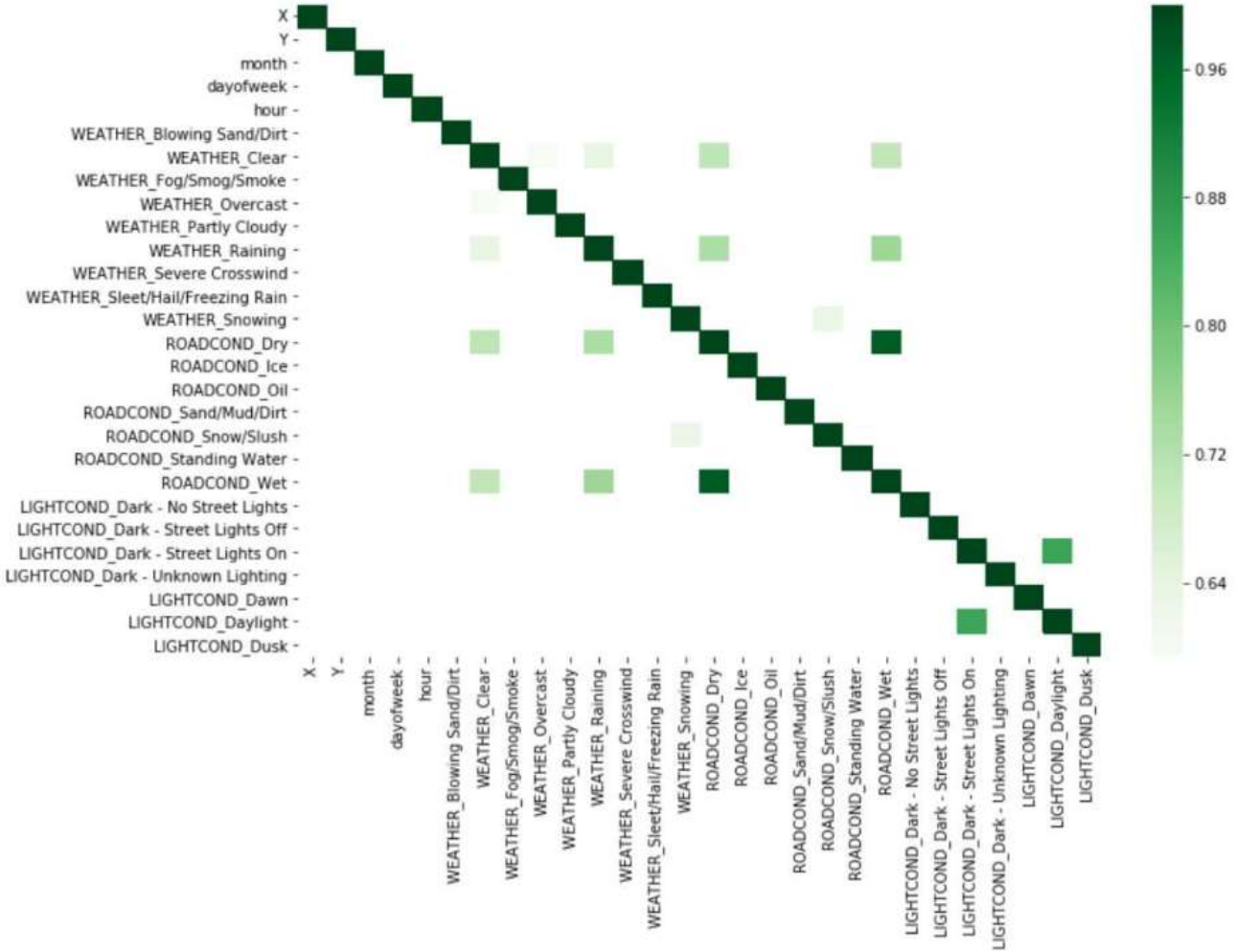
Extract month, day of week, and hour from variable ‘INCDTTM’.

Convert categorical to dummy variables. Remove ‘Other’ values.

Remove highly correlated variables

Standardize the data

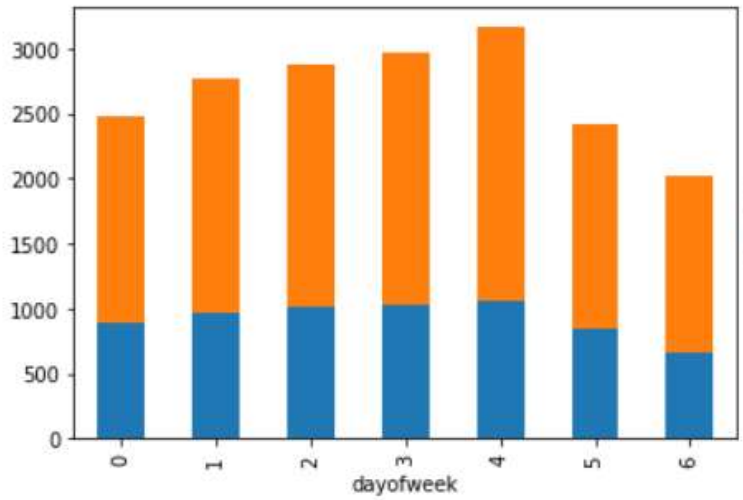
DATA PRE-PROCESSING

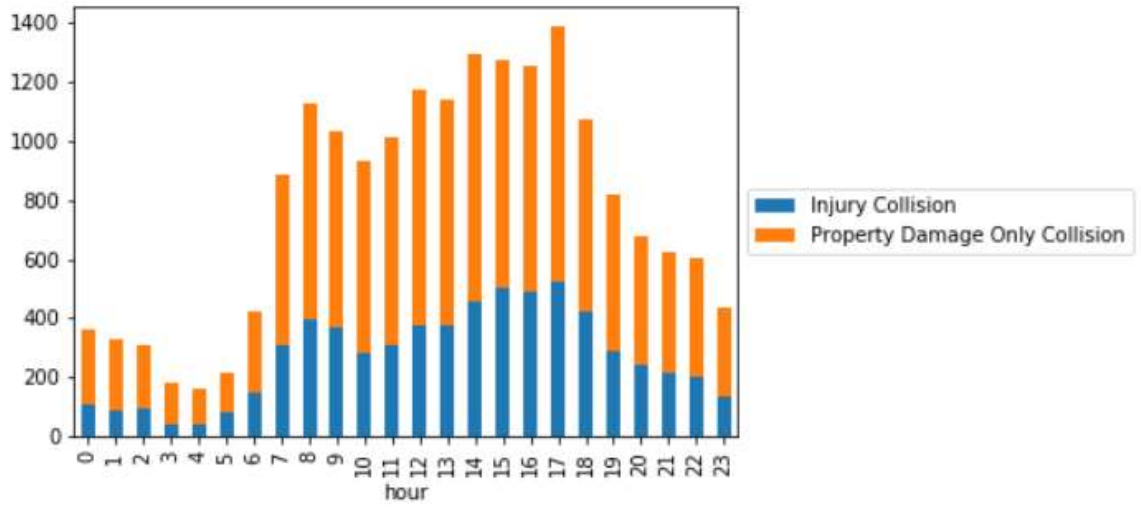


- 9 pairs of variables with high correlation
- Remove 6 variables:
  - WEATHER_Overcast
  - WEATHER_Raining
  - ROADCOND_Dry
  - ROADCOND_Wet
  - ROADCOND_Snow/Slush
  - LIGHTCOND_Dark - Street Lights On

DATA EXPLORATORY

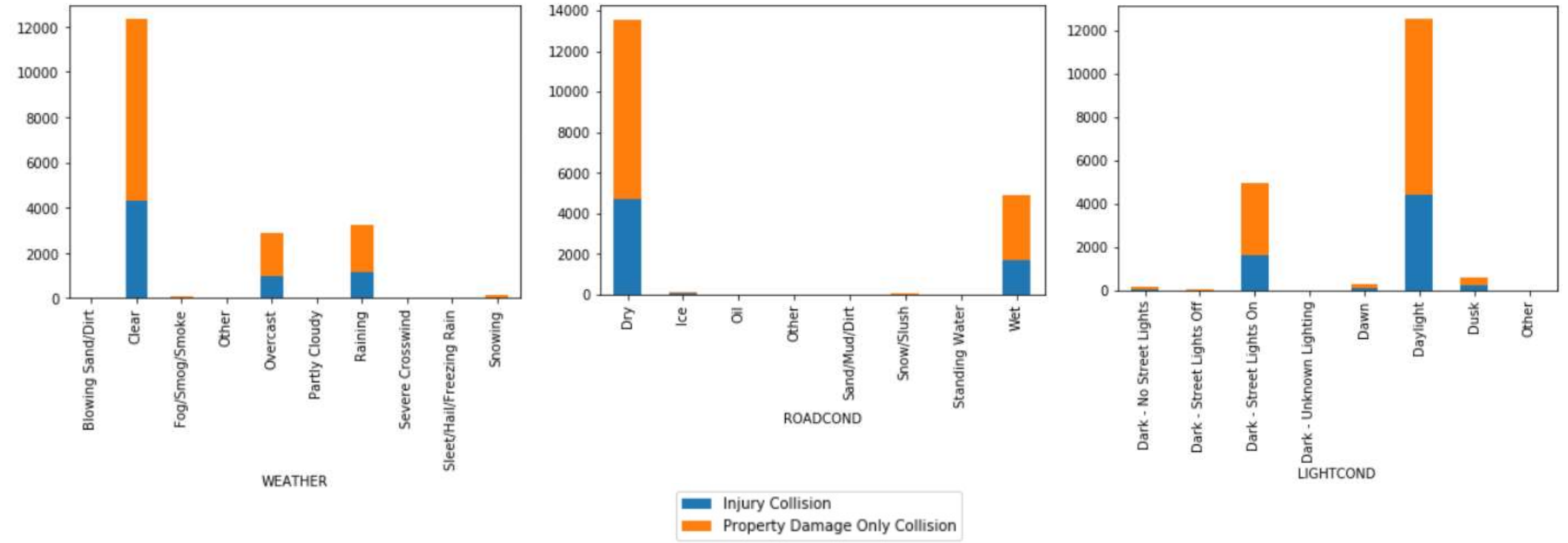
- About two thirds of collisions are involved with property damage only while injuries incur in the remaining
- Accidents are more likely to happen during weekdays than weekends, especially on Fridays
- Collisions are more likely to happen between 7am to 7pm – the time when people need to travel to work and school, especially afternoon peak hours





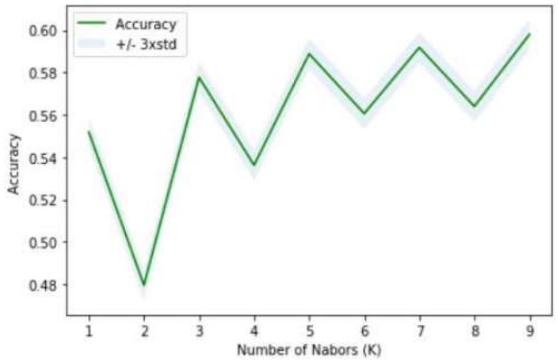
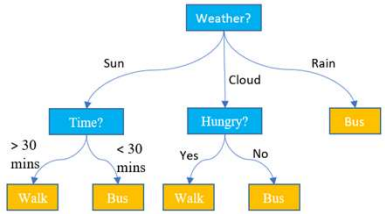
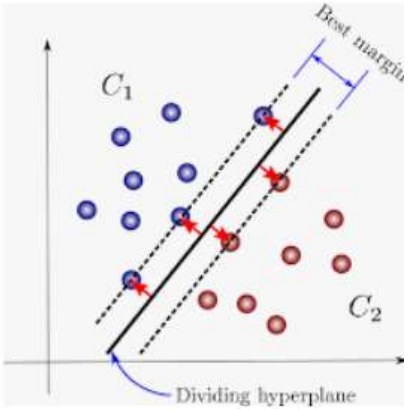
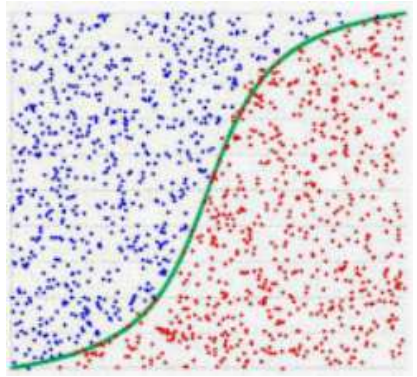
DATA EXPLORATORY

- 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.



MODELLING

- Try 4 models: kNN, Decision Tree, Support Vector Machine, and Logistic Regression to find out the best model

kNN	Decision Tree	SVM	Logistic Regression
			
k = 9	max depth = 9	sigmoid kernel	regularization = 0.01

CONCLUSION

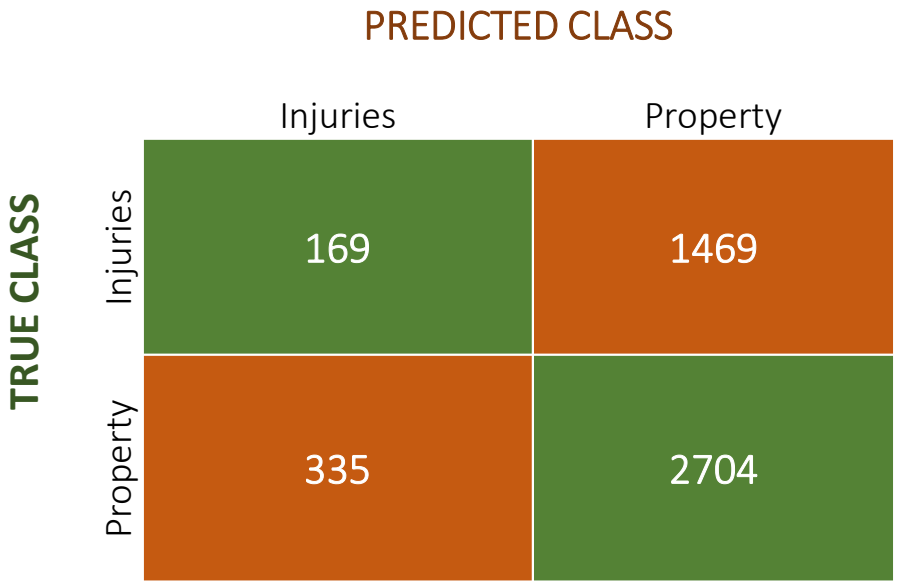
- Performance summary of the selected classification models. Best performer per criterion is labelled in red.
- kNN and Logistic Regression each takes a lead in half of the criteria.
- Logistic regression model failed to predict accidents with injuries.
- kNN has overfitting issue with low accuracy level on test set.
- SVM is the best choice. SVM is better at predicting accidents with injuries - 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

CONCLUSION

- Performance summary of SVM model
- Train using sigmoid kernel

Accuracy on train set	61%
Accuracy on test set	61.4%



RECOMMENDATION

The model can become a built-in function of GPS devices to give warnings to drivers.

Drivers can learn about places with high likelihood of accidents on their planned route and find ways to reroute if possible.

LIMITATION

Low accuracy level of 61%. Possible reasons: small data size, lack of important variables to cause accidents as carelessness, speeding, etc.

Data includes collisions in Seattle only with inputs being Seattle locations – model cannot be used outside Seattle.

Only two classes in the target variable – limits the prediction capabilities.


Thank
you!!!