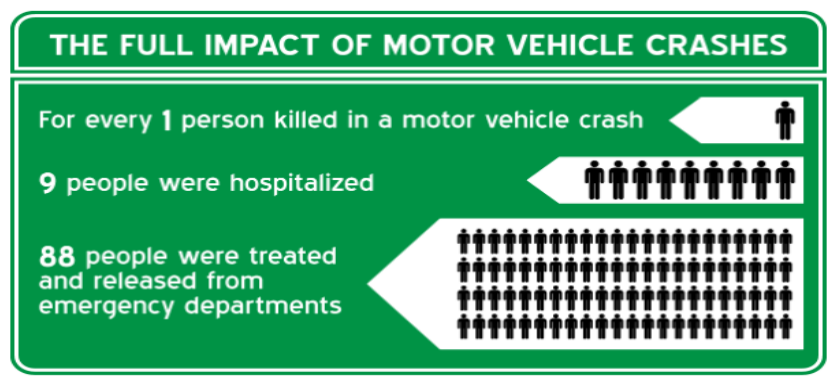
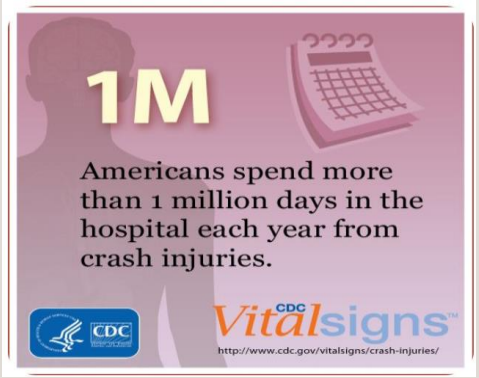
MACHINE LEARNING TO PREDICT ACCIDENT SEVERITY

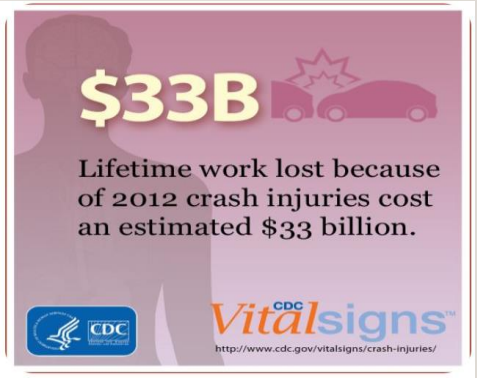
JOHANNES DANUSANTOSO

BACKGROUND

- Traffic accidents are a major cause of death in the US
- Traffic accidents (injury and non-injury) lead to huge total economic cost and
  - decreased quality of life
  - the immeasurable burden on the victims' families and friends.







OBJECTIVE

Predict the severity of traffic accidents using Machine Learning with Python

TARGET AUDIENCE

- Public traffic and safety officials
- General population

DATA SOURCE

- We use the collision dataset published by the City of Seattle.
- It is a public data and has more than 200k data points from 2004 to present with 40 attributes.

DATA UNDERSTANDING

- Group the target to 2 classes:
  1. Property damage only (no injuries)
  2. Injuries (including fatalities)
- Imbalanced datasets: ~70% of the data belong to class 1

DATA PREPARATION

- Transform the categorical values to numerical values.
- Perform attribute selection using Univariate Feature Selection and determine the top attributes contributing to model accuracy:

WEATHER: description of the weather conditions during the time of the collision.

Original Values	*New Values*	*Frequency*
Clear	1	114807
Raining	2	34038
Overcast	3	28556
blank		26641
Unknown	99	15131
Snowing	4	919
Other	99	860
Fog/Smog/Smoke	5	577
Sleet/Hail/Freezing Rain	6	116
Blowing Sand/Dirt	7	56
Severe Crosswind	8	26
Partly Cloudy	9	10
Blowing Snow	10	1

LIGHTCOND: the light condition during the collision

Original Values	*New Values*	*Frequency*
Daylight	1	119555
Dark - Street Lights On	2	50139
blank		26730
Unknown	99	13533
Dusk	3	6085
Dawn	4	2609
Dark - No Street Lights	5	1580
Dark - Street Lights Off	6	1239
Other	99	244
Dark - Unknown Lighting	7	24

ROADCOND: the condition of the road during the collision

Original Values	*New Values*	*Frequency*
Dry	1	128660
Wet	2	48737
(blank)		26560
Unknown	99	15139
Ice	3	1232
Snow/Slush	4	1014
Other	99	136
Standing Water	5	119
Sand/Mud/Dirt	6	77
Oil	7	64

ADDRTYPE: collision address type

Original Values	*New Values*	*Frequency*
Alley	0	879
Block	1	145118
Intersection	2	72027
blank		3714

MODELING

- Handle imbalance datasets
  - Balanced Bagging Classifier
  - Undersampling
- Train/test dataset: 80% / 20%
- Machine Learning algorithm
  - Support Vector Machine (SVM)
  - K-Nearest Neighbor
  - Decision Tree
  - Logistic Regression

1 = perfect accuracy for both F1-score and Jaccard index

RESULTS

Algorithm	F1-score		Jaccard index	
	BBC	UND	BBC	UND
SVM	0.650	0.595	0.494	0.425
KNN	0.585	0.569	0.469	0.398
Decision Tree	0.649	0.594	0.492	0.424
Log Regression	0.650	0.593	0.494	0.424

- Balanced Bagging Classifier (BBC) outperforms Undersampling (UND)
- SVM, Decision Tree, and Log Regression performs equally well

CONCLUSION

- We achieve ~65% accuracy in predicting the accident severity using SVM/Decision Tree/Log Regression with Balanced Bagging Classifier with the weather/road/lighting condition and the location of the accident as the input.
- Further study should be conducted to improve the model accuracy using different method of handling imbalance data.