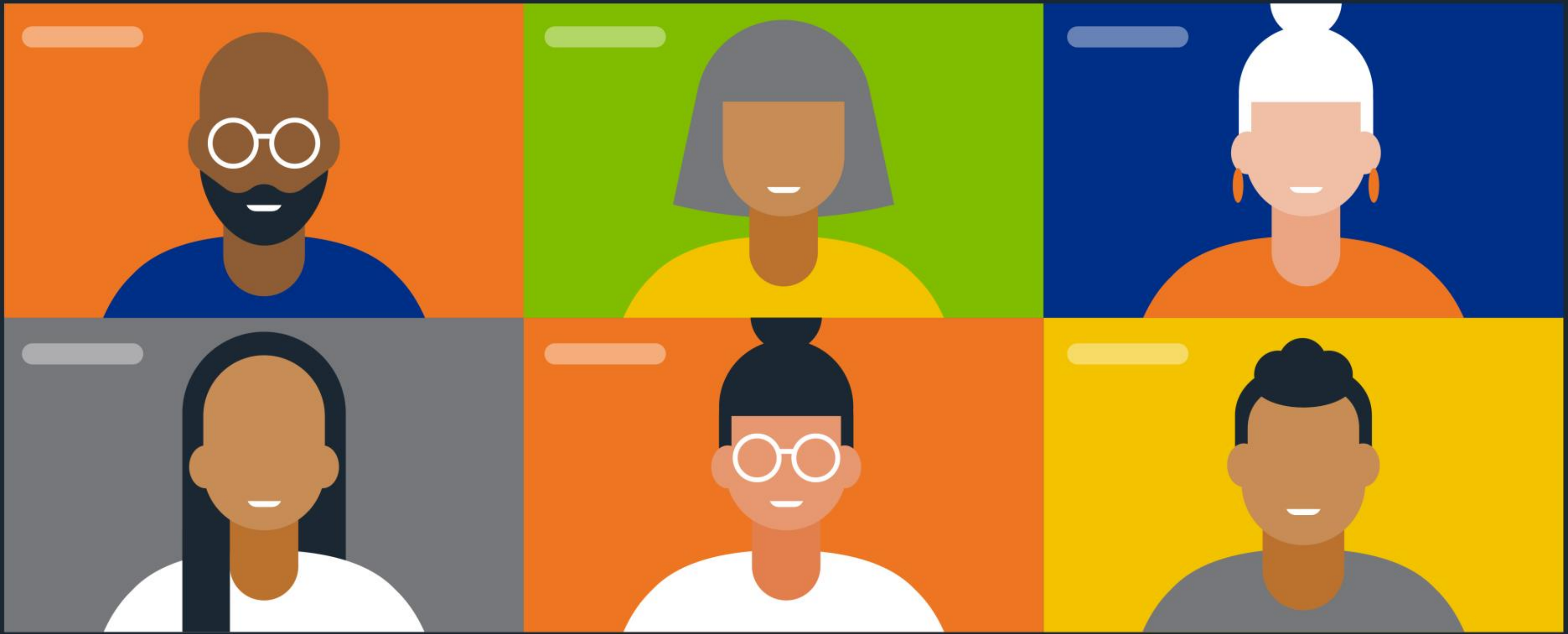


informs ANNUAL MEETING | 2020 VIRTUAL

Determinants of Car Accident Severity

Zejian Wu, Shuyu (Jade) Zhang, Jiayu Fan, Chenhao You, Yue Gao, Clark Univeristy

Source: Public Domain Pictures

Damage of Car Accident

36,560 people were killed in traffic crashes in 2018

- 1,038 children
- 9,378 speeding-related
- 4,985 motorcycle fatalities
- 6,283 pedestrians died
- 857 bicyclist deaths
- 885 large-truck occupants died

Over the past 10 years, the number of traffic deaths in urban areas has increased

- pedestrian deaths are up 69%
- bicyclist fatalities increased 48%
- motorcycle deaths are up 33%

United States Department of Transportation

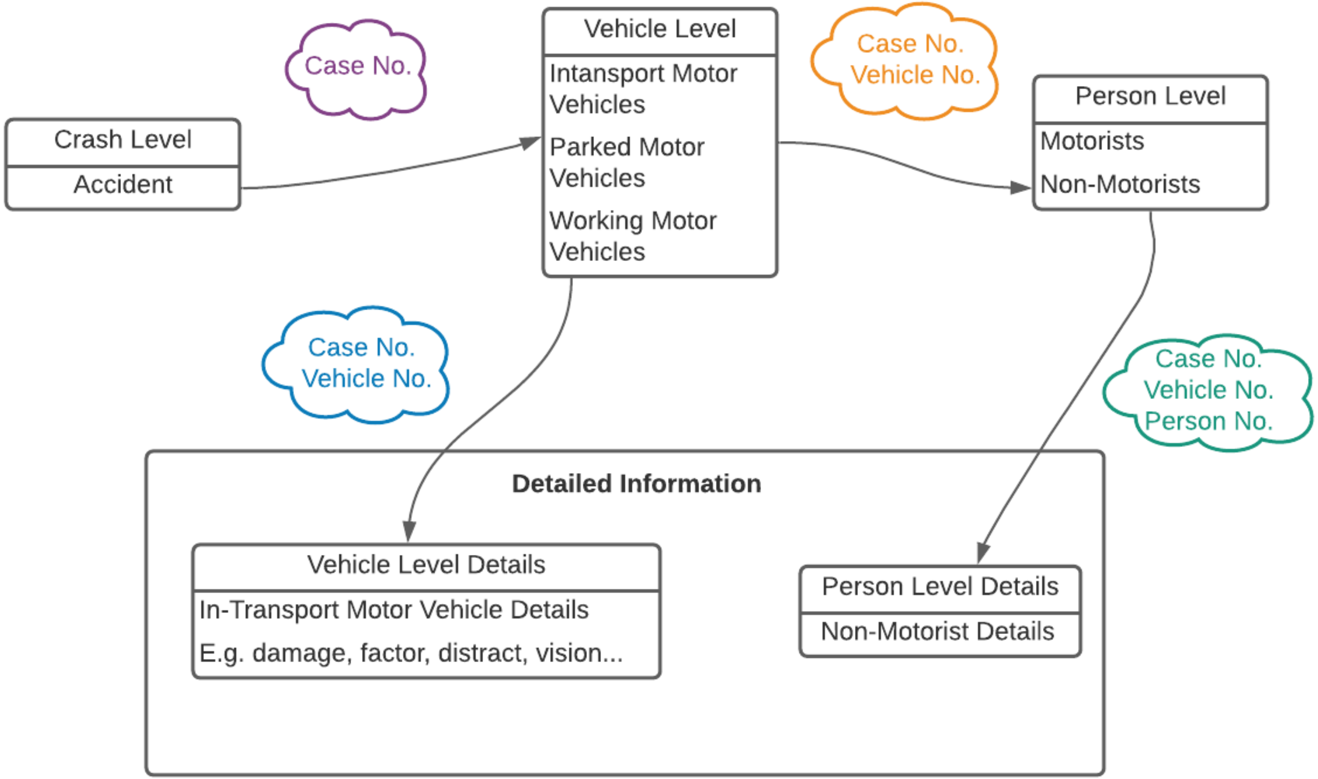
Basic Ideas

- Investigate the determinants of car accident
- Combined the information of vehicle, involved individuals, and surrounding environment
- 2018 car accident statistics from Crash Report Sampling System (U.S. Department of Transportation)

- Person Level & Vehicle Level
- Basic Machine Learning methods
- Make contribution to car accident and injury prevention

informs

Data Description



Crash Level

Accident

Case No.

Vehicle Level

Intansport Motor Vehicles

Parked Motor Vehicles

Working Motor Vehicles

Case No.
Vehicle No.

Person Level

Motorists

Non-Motorists

Case No.
Vehicle No.

Case No.
Vehicle No.
Person No.

Detailed Information

Vehicle Level Details

In-Transport Motor Vehicle Details

E.g. damage, factor, distract, vision...

Person Level Details

Non-Motorist Details

informs

Targets

1. Vehicle Level (N=51,260)
   - MXVSEV_IM

     Maximum Injury Severity in Vehicle

     Group 1: No Apparent Injury

     Group 2: Injury
   - MAXSEV_IM

     Maximum Injury Severity in Crash

     Group 1:No Injury of all Persons

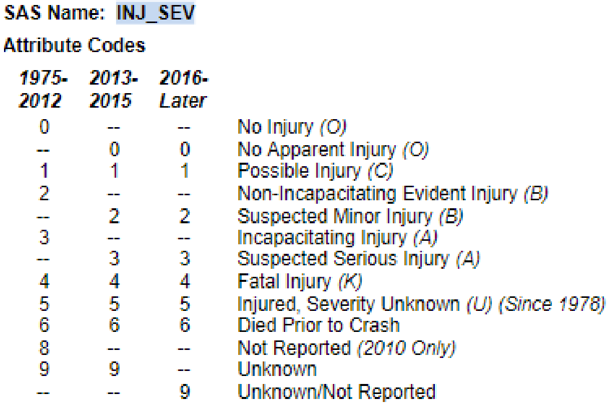
     Group 2:Injury

2. Person Level (N=62,933)
   - INJSEV_IM

     Maximum Injury Severity of Person

     Group 1:No or No Apparent Injury

     Group 2: Injury



SAS Name: INJ_SEV

Attribute Codes

1975-2012	*2013-2015*	*2016-Later*	
0	--	--	No Injury *(O)*
--	0	0	No Apparent Injury *(O)*
1	1	1	Possible Injury *(C)*
2	--	--	Non-Incapacitating Evident Injury *(B)*
--	2	2	Suspected Minor Injury *(B)*
3	--	--	Incapacitating Injury *(A)*
--	3	3	Suspected Serious Injury *(A)*
4	4	4	Fatal Injury *(K)*
5	5	5	Injured, Severity Unknown *(U) (Since 1978)*
6	6	6	Died Prior to Crash
8	--	--	Not Reported *(2010 Only)*
9	9	--	Unknown
--	--	9	Unknown/Not Reported

Attribute Codes for INJ_SEV

General View of Targets

1. Vehicle Level

- Maximum Injury Severity in Vehicle

Mean	Medium	Maximum	Minimum	Standard Dev
0.5503	0	5	0	0.9483

- Maximum Injury Severity in Crash

Mean	Medium	Maximum	Minimum	Standard Dev
0.9180	0	6	0	1.1226

2. Person Level

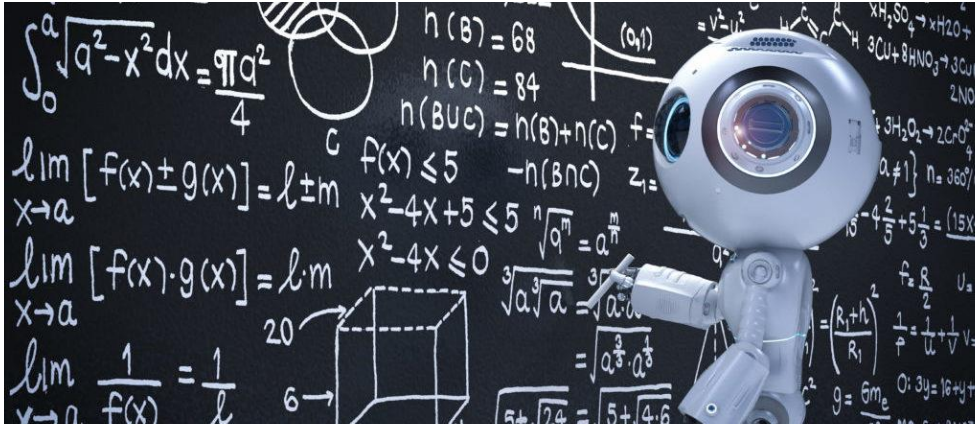
- Maximum Injury Severity of Person

Mean	Medium	Maximum	Minimum	Standard Dev
0.4926	0	6	0	0.9016

Algorithms

- Logistic Regression
- Support Vector Machine

- K-Nearest Neighbors
- Decision Tree
- Random Forest



$\int_0^a \sqrt{a^2-x^2}dx = \frac{\pi a^2}{4}$
$n(B) = 68$
$n(C) = 84$
$n(B \cup C) = n(B)+n(C) - n(B \cap C)$
$\lim_{x \to a} [f(x) \pm g(x)] = l \pm m$
$\lim_{x \to a} [f(x) \cdot g(x)] = l \cdot m$
$f(x) \leq 5$
$x^2-4x+5 \leq 5$
$x^2-4x \leq 0$
$(0,1)$

Source: Information Age Automating - data science and machine learning for business insights

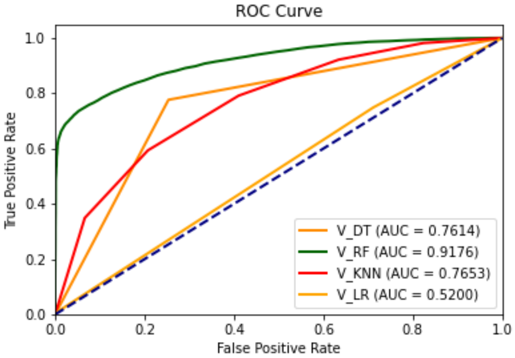
Results - Performance

Vehicle Level

1. Maximum Injury Severity in Vehicle

	Logistic Regression	SVM	KNN	Decision Tree	Random Forest
Accuracy	49.81%	48.52%	69%	81.17%	**84.15%**
AUC	0.5200	0.5000	0.7652	0.7614	**0.9176**

Best ROC CURVE – Random Forest



Confusion Matrix – Random Forest

	Predict: No Injury	Predict: Injury	Sensitivity	Specificity
Actual: No Injury	9812	808	75.84%	92.39%
Actual: Injury	2547	7994		

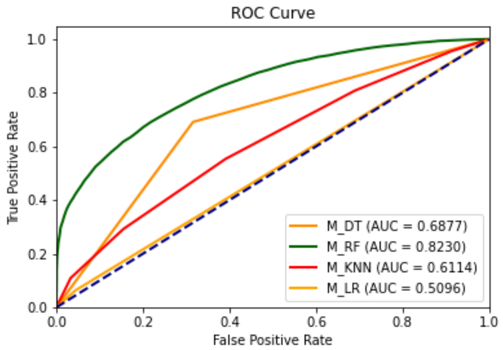
Result - Performance

Vehicle Level

2. Maximum Injury Severity in Crash

	Logistic Regression	SVM	KNN	Decision Tree	Random Forest
Accuracy	49.63%	49.81%	58.1%	68.77%	**73.59%**
AUC	0.5096	0.5000	0.6114	0.6877	**0.8230**

Best ROC CURVE – Random Forest



Confusion Matrix – Random Forest

	Predict: No Injury	Predict: Injury	Sensitivity	Specificity
Actual: No Injury	5972	1660	69%	78.25%
Actual: Injury	2401	5345		

informs

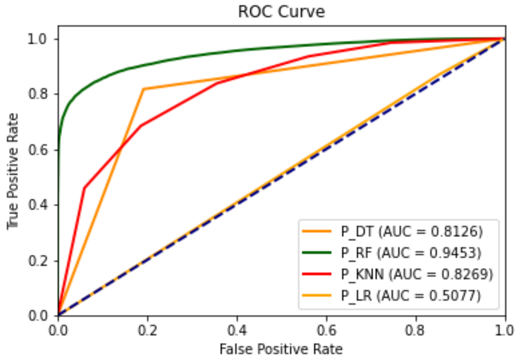
Result - Performance

Person Level

1. Maximum Injury Severity of Person

	Logistic Regression	SVM	KNN	Decision Tree	Random Forest
Accuracy	49.77%	49.77%	74.05%	81.17%	**87.99%**
AUC	0.5077	0.5000	0.8269	0.8126	**0.9453**

Best ROC CURVE – Random Forest



Confusion Matrix – Random Forest

	Predict: No Injury	Predict: Injury	Sensitivity	Specificity
Actual: No Injury	12620	951	82.95%	92.99%
Actual: Injury	2293	11154		

informs

Result – Feature Importance

Vehicle Level

1. Maximum Injury Severity in Vehicle

Variable	Description
NUMOCCS	a count of the number of occupants in this vehicle
ACC_TYPE	the type of crash this vehicle was involved in
BDYTYP_IM	a classification of this vehicle based on its general body configuration, size, shape, doors, etc.
MODEL	the model of this vehicle within a given make

Result – Feature Importance

Vehicle Level

2. Maximum Injury Severity in Crash

Variable	Description
ACC_TYPE	the type of crash this vehicle was involved in
MINUTE_IM	the minutes after the hour at which the crash occurred
MAK_MOD	the 5-digit combination of two data elements ("Vehicle Make” code (MAKE) followed by the 3-digit “Vehicle Model” code (MODEL))
MODEL	the model of this vehicle within a given make

informs

Result – Feature Importance

Person Level

1. Maximum Injury Severity of Person

Variable	Description
AIR_BAG	air bag availability and deployment for this person
SEX_IM	the sex of this person involved in the crash
AGE_IM	the age of this person involved in the crash
MINUTE_IM	the minutes after the hour at which the crash occurred

Conclusion

1. Random Forest and Decision Tree obtained the most satisfying results
2. K-Nearest Neighbors, Logistic Regression, and Support Vector Machine cannot provide good estimation for this problem
3. At vehicle-level, a count of the number of occupants in this vehicle, the type of crash this vehicle was involved in, and the make and model of the vehicle play important roles in accident severity
4. At person-level, sex, age, and air bag availability and deployment are the most important indicators.

Future Research

1. Integrate more data (before and after 2018) at a more detailed-level
2. Alternative machine learning algorithms: Neural Networks, Naïve Bayes Classifier
3. More feature engineering: logarithm transformation (age/size), Grouping (model/make), and Categorical Imputation.
4. Investigate further relationship among the features

zewu@clarku.edu

Thank you!