Predicting Car Accident Severity

1. Introduction

Background and Problem

No one anticipates being involved in a car accident

“

Car accidents are the cause of death for 1.35 million people per year globally and cause of 3% of gross domestic product loss in most countries.

World Health Organization

Predicting accidents severity is valuable for city management

- Car accidents cost most countries 3% of their GDP.
- Car accidents injuries can be prevented.
- City management is interested in predicting accidents severity and lowering risks of severe accidents.

The results of this project could be used to adjust construction plans for road infrastructure and bridges, to impose traffic restrictions in the city center, to plan speed limits at particular road parts, etc.

Businesses benefit from accidents prediction

- Banks and property insurance companies.
- Car dealers, car parts suppliers, repair workshops.
- Health insurance and medical organizations.
- Mass media.

Individual drivers and general public could also be interested in results of this project, as people may take extra precaution measures in particular parts of roads, or types of junctions, or during specific weather.

2. Data

Acquisition and cleaning

Data acquired via City of Seattle open data platform

Dataset

The CSV file of the weekly updated collisions dataset is available here - link.

Timeframe: 2004 to Present.

Shape: 221783 rows, 40 columns.

Metadata

The PDF file of the description of collisions attributes and codes is available here - link.

Data cleaning

- Duplicate columns, empty rows, highly similar features and useless report codes were dropped.
- Date and time features were dropped after extracting month and year data.
- Categorical values were replaced with numerical codes.
- Data type mismatch was fixed.

Cleaned dataset contains 169,089 samples and 22 features.

169,089 samples

in the cleaned dataset

22 features

18 were used for 5 ML models

3 classes

target variable

135,271 samples
in the train set

33,818 samples
in the test set

100%
Final accuracy

Target value - accident severity code

Class 1

Severity code: 1

Meaning: property damage only

Color: gold

Class 2

Severity code: 2

Meaning: injuries

Color: blue

Class 3

Severity code: 2b, 3

Meaning: serious injuries and fatalities

Color: red

3. Methodology

Libraries and techniques

Planning analysis steps and methods

- Load, read and inspect the dataset.
- Fix missing data and type mismatch.
- Investigate relations between accidents severity and attributes.
- Pick relevant attributes and build ML models.
- Assess and compare models' performance.
- Improve models and describe the final model with recommendations.

Loading the libraries

- Numpy - to work with arrays
- Pandas - to work with tabular data
- Datetime - to extract the data from timestamp
- Plotly, Matplotlib, Seaborn - to plot the data
- Scikit-learn - to build and assess ML models

Picking machine learning techniques for classification problem

- Decision Tree - to be able to process the influence of multiple different attributes.
- Random Forest - to improve the performance of Decision Tree model (if necessary).
- Logistic Regression - to model a nonlinear association.
- Naive Bayes - to use in multi class prediction.
- k-Nearest Neighbors - to check how good it would perform in out case study.

4. Analysis

Data exploration and hypothesis check

Many factors contribute to severity of accidents

- Environment conditions (e.g., weather).
- Driver's behavior (e.g., driving rules violation).
- Participants (e.g., car, pedestrian or cyclist).
- Specific road situation (see SDOT_COLCODE).

Accidents severity is high if environment conditions are good

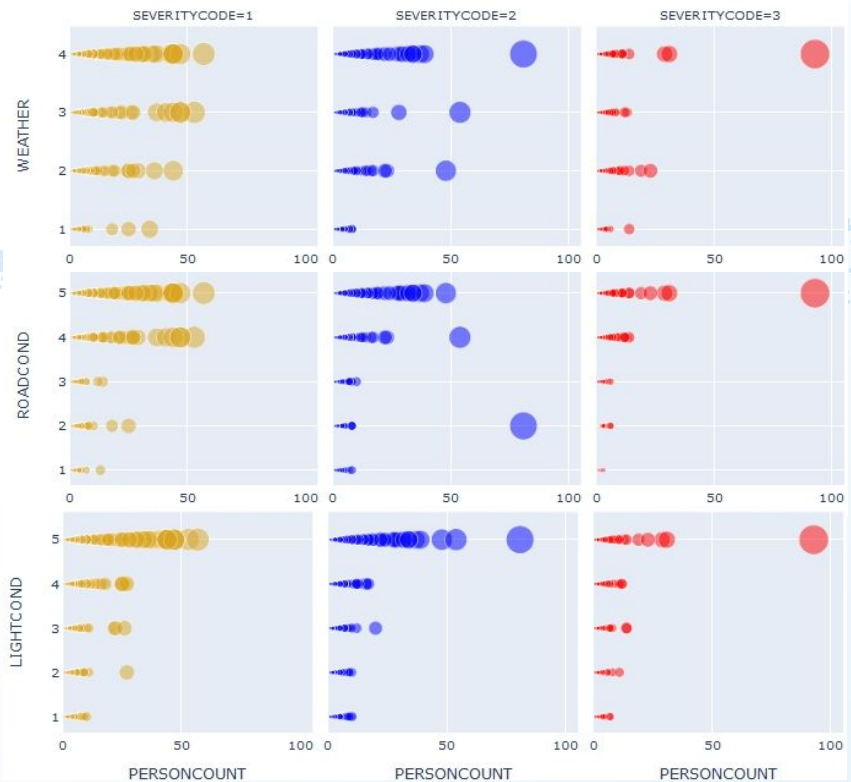
Model	Number of cases in each severity class, %			Dataset summary		
	1	2	3	Number of cases, %	People involved, %	Vehicles involved, %
Clear weather	**55.3**	**61.8**	**64.9**	**57.4**	**58.7**	**59.0**
Dry road	**62.1**	**69.1**	**72.5**	**64.3**	**65.7**	**66.1**
Daylight	**57.0**	**66.1**	**56.5**	**59.8**	**61.2**	**61.8**

Accidents are more likely to be severe in good weather, on a dry road, in a light of a day.

Accident severity, environment conditions & people involved

Clear weather and good road conditions contribute the most to number of accidents and number of people involved.

Probably, road traffic is more dense in such conditions, when most drivers do not expect accidents to happen.



Accidents severity grows if driver violates the rules

Model	Number of cases in each severity class, %			Dataset summary		
	1	2	3	Number of cases, %	People involved, %	Vehicles involved, %
Inattention	14.1	17.7	10.9	15.1	16.6	16.0
Driving under the influence	4.0	6.1	14.6	4.8	5.1	4.8
Pedestrian not granted right of way	0.3	7.2	13.9	2.6	2.4	1.4
Speeding	4.3	6.1	13.5	5.0	5.0	1.4
Not one violation	74.5	65.5	4.4	71.5	70.5	71.7

Majority of accidents happen without direct violation of rules

The dataset could be incomplete in terms of possible causes of collision (data limitation)

Inattention: 16% of cases (0.2% of class 3).

DUI: 5% of cases (0.2% of class 3)

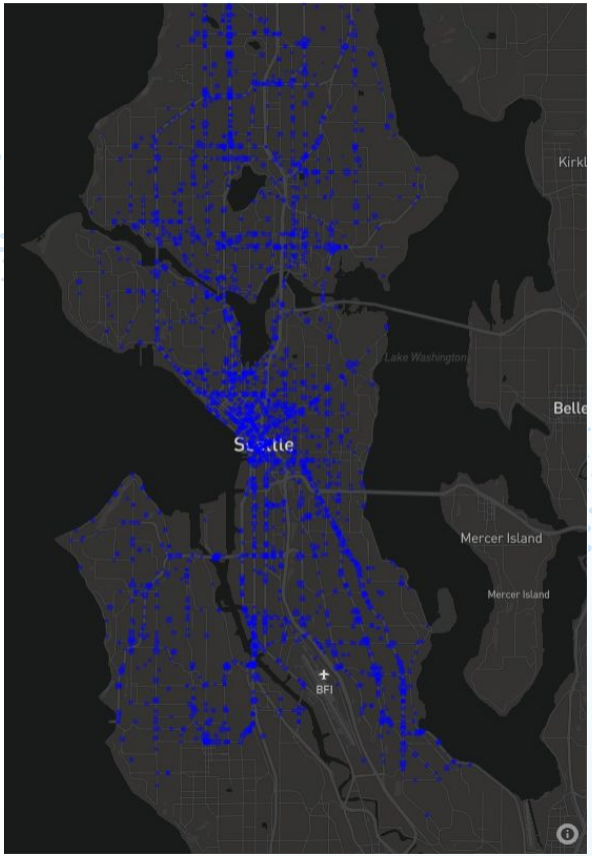
Not granting right of the way to pedestrian: 2.4% of cases (0.2% of class 3)

Accidents severity grows if pedestrians or cyclists involved

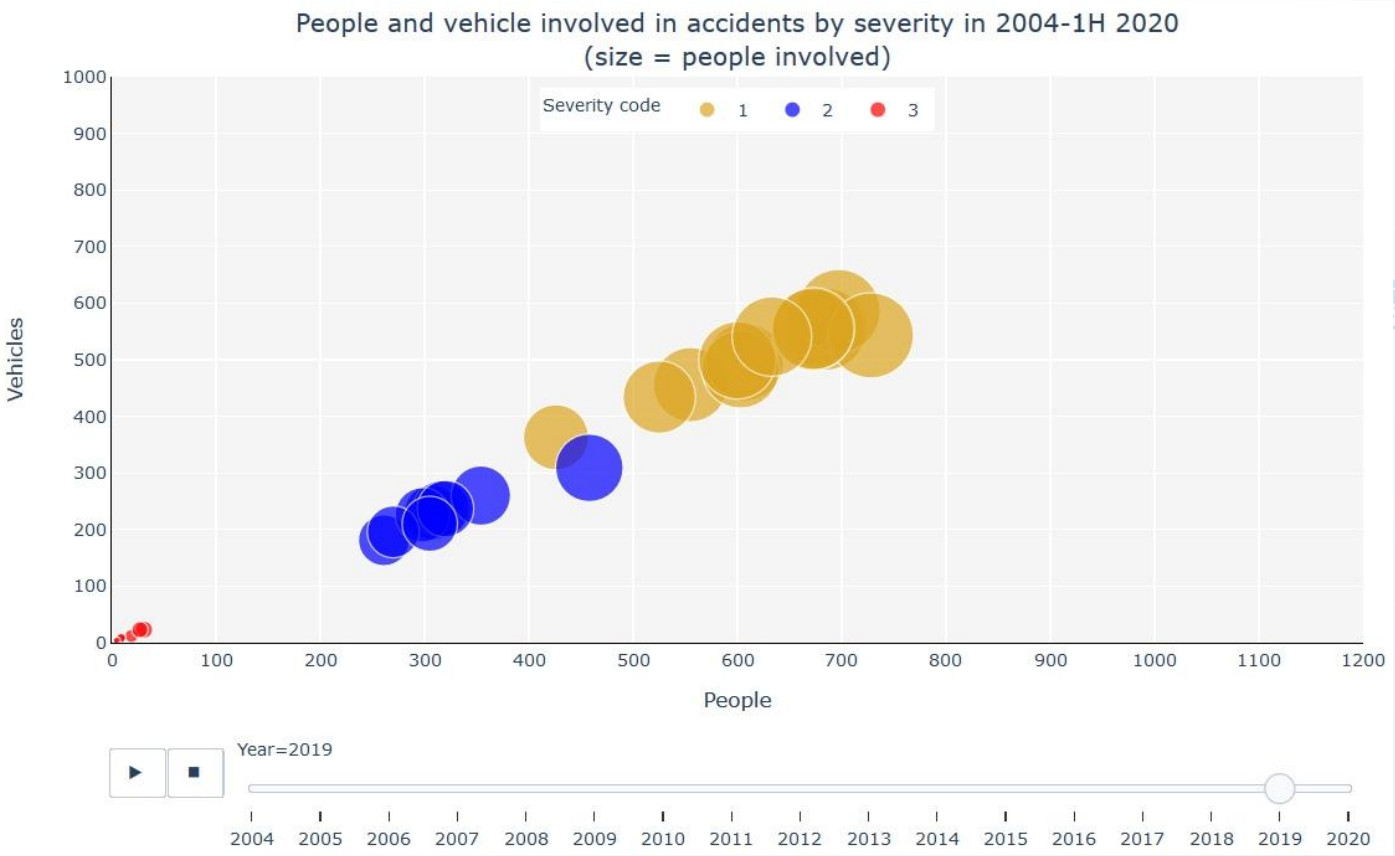
Model	Number of cases in each severity class, %			Dataset summary		
	1	2	3	Number of cases, %	People involved, %	Vehicles involved, %
No cyclists & pedestrians	**99.0**	**81.1**	**56.8**	**93.0**	**93.7**	**96.3**

Car drivers have protection in case of collisions, but pedestrians and cyclists are extremely vulnerable. If pedestrian or cyclist is involved, most likely they would be seriously injured or killed in the accident.

Accident frequency is higher in central city locations in 2019



People & vehicles involved in accidents, by month & severity



5. Modeling

Build, assess, compare and improve

Building and assessing 5 models

- Decision tree.
- Random forest.
- Naive Bayes.
- Logistic Regression.
- kNN.

Models performance comparison: Average vs Weighted

Model	Precision		Recall		F1 Score	
	Macro average	Weighted average	Macro average	Weighted average	Macro average	Weighted average
Decision Tree	1.00	1.00	1.00	1.00	1.00	1.00
Random Forest	1.00	1.00	1.00	1.00	1.00	1.00
Naive Bayes	0.96	0.99	0.99	0.99	0.98	0.99
Logistic Regression	0.99	0.99	0.89	0.99	0.93	0.99
k-Nearest Neighbors	0.95	0.92	0.61	0.92	0.61	0.91

Weights of classes improved overall ML models performance

- Decision tree & Random forest: good without weights.
- Naive Bayes: F1 score +1% and precision +3%.
- Logistic Regression: F1 score +6% and recall +10%.
- kNN: F1 score +30% and recall +31%, but -3% in precision.

So far decision tree model is the best, and kNN - the worst.

Decision tree / Random forest performed the best - result #1

	Precision	Recall	F1 score	Accuracy
Class 1	1.00	1.00	1.00	1.00
Class 2	1.00	1.00	1.00	1.00
Class 3	1.00	1.00	1.00	1.00

Naive Bayes model performance is the result #2

	Precision	Recall	F1 score	Accuracy
Class 1	1.00	0.99	0.99	0.99
Class 2	0.98	1.00	0.99	0.99
Class 3	0.92	1.00	0.96	1.00

Logistic Regression model performance is the result #3

	Precision	Recall	F1 score	Accuracy
Class 1	1.00	1.00	1.00	1.00
Class 2	0.99	1.00	0.99	0.99
Class 3	1.00	0.68	0.81	0.81

kNN model performance is the worst - result #4

	Precision	Recall	F1 score	Accuracy
Class 1	0.91	0.99	0.95	0.92
Class 2	0.92	0.82	0.87	0.92
Class 3	1.00	0.01	0.03	0.92

Confusion matrices: Decision Tree and Naive Bayes

Decision Tree

	Class 1	Class 2	Class 3
Class 1	22114	0	0
Class 2	0	11103	0
Class 3	0	0	601

F1 Score: 100%

Naive Bayes

	Class 1	Class 2	Class 3
Class 1	21820	280	14
Class 2	0	11064	39
Class 3	0	0	601

F1 Score: 96-99%

Confusion matrices: Logistic Regression and kNN

Logistic Regression

	Class 1	Class 2	Class 3
Class 1	22114	0	0
Class 2	0	11103	0
Class 3	48	145	408

F1 Score: 81-100%

kNN

	Class 1	Class 2	Class 3
Class 1	21847	267	0
Class 2	1994	9109	0
Class 3	0	494	8

F1 Score: 3-95%

6. Conclusion

Results

Results

I analyzed the relationship between accidents severity and features that influence on accidents severity. I identified environment, driver's behavior, location, participants and crash circumstances among the most important features.

Each feature alone is not enough to predict the outcome of an accident, so they were used together to get good prediction results.

Decision tree classification model demonstrated the best prediction result of all models that were built.

7. Discussion

Recommendations and future development

Recommendations to discuss

1. Create and publish car accidents risks map, based on accidents number, number of people involved and accidents severity. The map would indicate the streets where drivers should be extra cautious being used as a part of GPS navigation software for car drivers, providing recommendations with consideration of current and forecasted weather.
2. Convert city center streets from the motor vehicles traffic zones to bike lanes and walking zones, like in Stockholm or Copenhagen.
3. Use the car accidents risk map to adjust existing speed limits in Seattle, and increase fines for speeding (now they are about $70).
4. Increase fines for distracted driving (Driving Under the Influence of Electronics), $140-$240 ticket that came into force in 2017. The fine should depend on the accident risk zone where the offence was registered.

Additional data for meaningful insights

1. Capture more accident-specific data on drivers demographics: age, sex, income level, marital status, education level, occupation, goal of the trip.
2. Add vehicle speed data at the time of accident to the dataset (now it only indicates whether there were violations of speed limits or not, but no data on what those speed limits are).
3. Capture data on city streets traffic to relate with the geolocations of accidents: daily traffic volumes and speed limitations in place, etc. It would allow comparison between neighborhoods that differ in traffic density.

Additional data would help develop more detailed recommendations on city construction planning and traffic management in the future.