PREDICTING ACCIDENT SEVERITY

Coursera - IBM Data Science
Antoine Rousseau

PROBLEM

Traffic accidents are an endemic problem in the US. In 2018 alone, 12 million vehicles were involved in a car crash. Individuals take the road, often out of necessity, not thinking twice about the conditions they will face, and how much they will put themselves and others at risk.

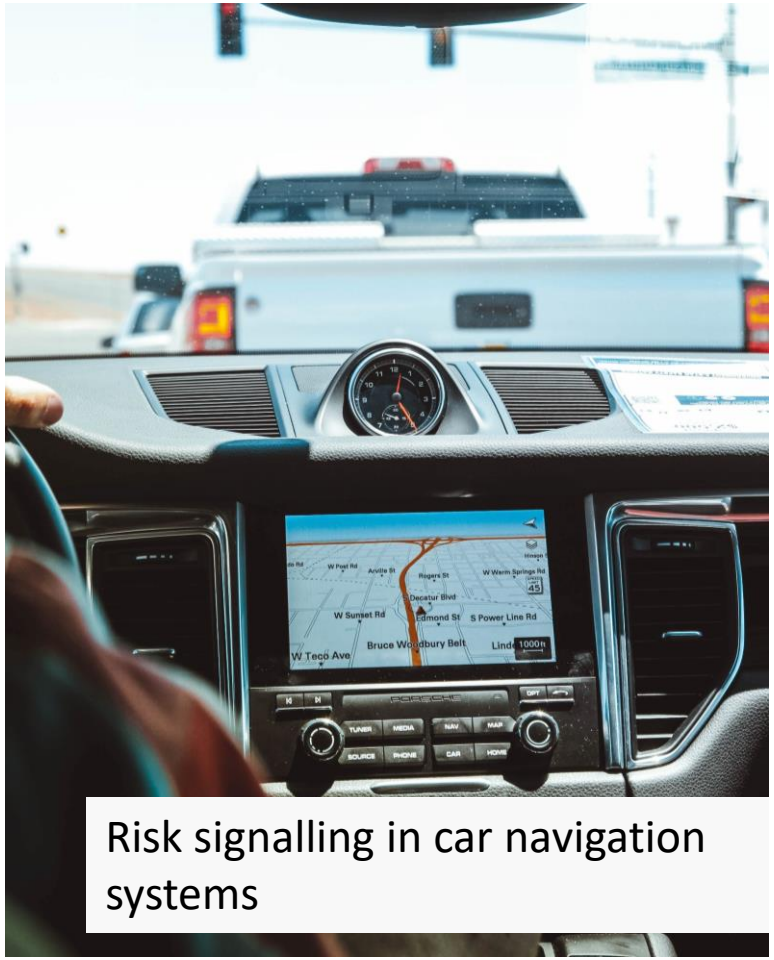
What if we could predict the severity of an accident, if one were to occur, based on time, date and environmental data, and thereby provide information that could reduce their frequency?

OUTCOME

AUTHORITIES

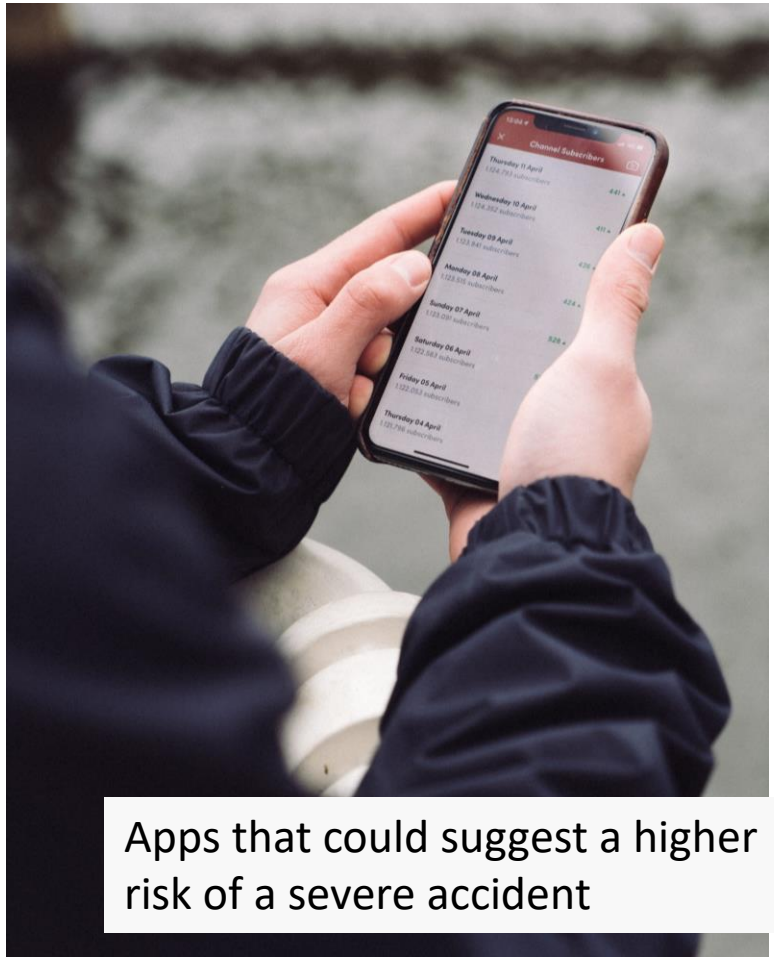
Optimization of rapid-response services

CAR MANUFACTURERS



Risk signalling in car navigation systems

INDIVIDUALS



Apps that could suggest a higher risk of a severe accident

METHODOLOGY

Our analysis is built upon the Seattle Collisions dataset. It contains, in tabular form, 31 attributes detailing each of the 194 673 collisions that took place in the Seattle from 2004 to today. Each entry is labelled according to severity.

1. We will start by describing the dataset, gauging the quality of the data, and assessing the relevance to the problem of the available attributes.
2. We will then transform it, selecting the useful attributes, addressing missing or unknown entries, and encoding it.
3. We will then run several classification algorithms and evaluate their performance.
4. Finally, we will discuss the findings and assess how the model answers the problem statement.

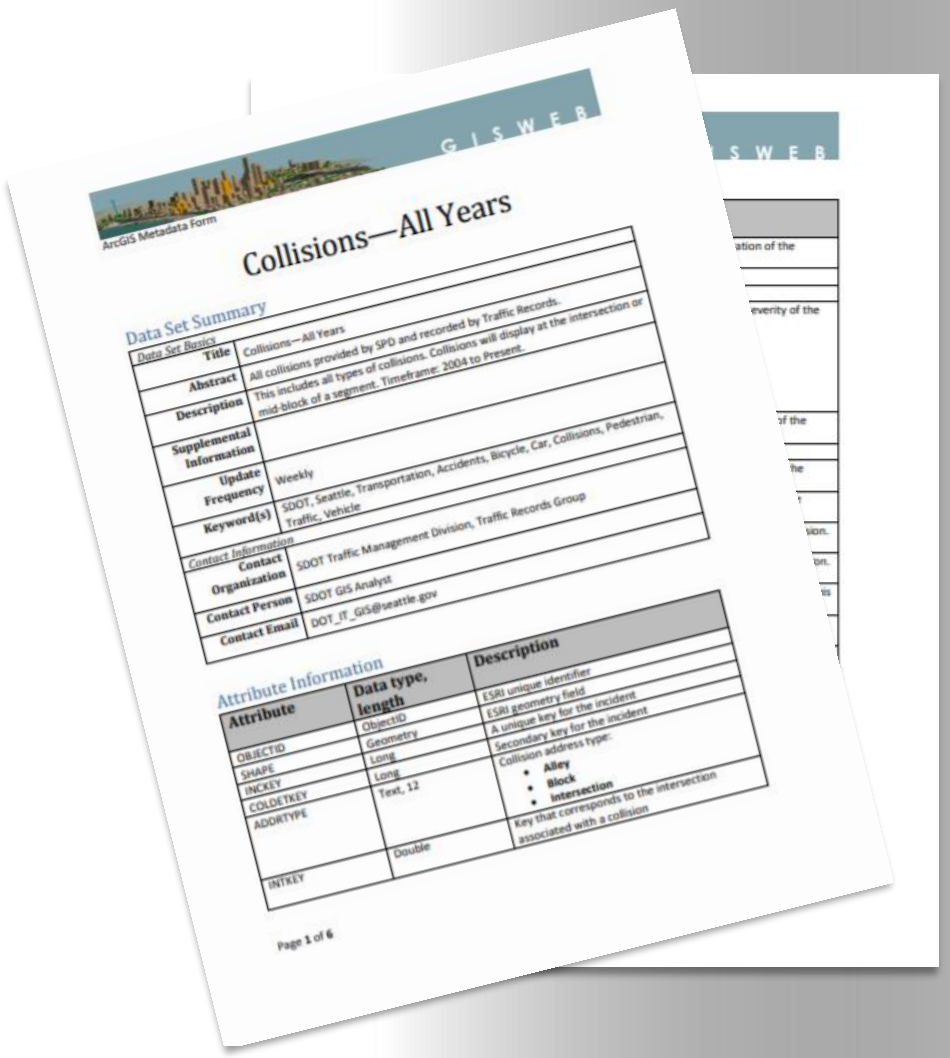
DATA UNDERSTANDING

The first step in analysing the data is to get an understanding of the attributes. Out of the 31 attributes only 5 are of interest to us to solve the problem:

- **Time/Date**: INCDATE, INCDTTM
- **Environmental**: WEATHER, ROADCOND, LIGHTCOND

Selecting these gives us the following data frame

	SEVERITYCODE	INCDATE	INCDTTM	WEATHER	ROADCOND	LIGHTCOND
0	2	2013/03/27 00:00:00+00	3/27/2013 2:54:00 PM	Overcast	Wet	Daylight
1	1	2006/12/20 00:00:00+00	12/20/2006 6:55:00 PM	Raining	Wet	Dark - Street Lights On
2	1	2004/11/18 00:00:00+00	11/18/2004 10:20:00 AM	Overcast	Dry	Daylight
3	1	2013/03/29 00:00:00+00	3/29/2013 9:26:00 AM	Clear	Dry	Daylight
4	2	2004/01/28 00:00:00+00	1/28/2004 8:04:00 AM	Raining	Wet	Daylight



DATA PREPARATION

Before we ingest the data frame into the different models, we perform the following operations:

- Change the types of all the attributes from Object to:
  - datetime for the Time/Date attributes
  - category for the Environmental attributes
- Change INCDTTM to the hour of the day (0-23)
- Change INCDATE to day of the week (0-6)
- Replace Unknown/Other entries by n/a's
- Fill n/a's with most frequent values

Further, numerical encoding results in the following data frame:

	SEVERITYCODE	INCDATE	INCDTTM	WEATHER	ROADCOND	LIGHTCOND
0	1	2	14	1	1	0
1	0	2	18	2	1	1
2	0	3	10	1	0	0
3	0	4	9	0	0	0
4	1	2	8	2	1	0

MODELLING & RESULTS

As this is a classification problem, we will compare the performance of some common machine learning algorithms: Logistic Regression (LR), KNN, and Decision Tree (DT). We train-test split our data, leaving 10% for testing. We also add a Dummy Classifier which will return the most frequent class.

The parameter choices are as follows:

- KNN: neighbours = 3
- LR: class-weight = balanced
- DT: max depth = 5, class-weight = balanced

Running the models gives the following scores:

	Accuracy	F1	Jaccard
Dummy Classifier	0.706030	0.000000	0.000000
Nearest Neighbors	0.618708	0.273039	0.158104
Logistic Regression	0.528354	0.395444	0.246451
Decision Tree	0.496302	0.436566	0.279236

DISCUSSION

The scores obtained for all three models are low. This indicates that further iterations are needed to improve the performance and conclude if our problem can be resolved with the current data set.

Alternative A: Removing rows that contain n/a, Unknown or Other data instead of filling them with the most frequent category. **Result** : this worsens the performance.

Alternative B: One-hot encoding features (n/a's still removed). **Result** : Once again this does not dramatically improve model performance.

At this stage it is worth asking whether the scores can be improved with the features selected. We should go back to the original problem and reshape it, in a perhaps less ambitious way, using a wider set of features.

CONCLUSION

Our original question, **“What if we could predict the severity of an accident, if one were to occur, based on time, date and environmental data, and thereby provide information that could reduce their frequency?”** was not answered in this study.

We were however able to identify that taking date, time and environmental data available in the Seattle dataset was insufficient to build a reliable model.