Accident Severity Predictability

Using a Range of Variables to Predict Severity of an Accident

Author: Elliot Eisenberg

IBM Data Science Capstone

What is the Issue?

Many factors contribute to the severity of an accident:

Number of cars	Number of pedestrians	Number of total people involved	Time of the day	The list goes on…

Given a large dataset of information, can we train a model to predict the severity of an incident based on a set of criteria?

Who Cares?

Emergency Responders:	Insurance Companies:	Legislators:	Drivers:	Navigation-Focused Companies:
• Knowing what level of response an incident requires, depending on the data	• Possibility of 'parametric' insurance policy, where meeting given criteria leads to coverage or not (time of day, driving conditions, number of passengers, etc.)	• Determining speed limits	• Understanding if there is heightened risk of driving under a set of circumstances.	• Providing driving recommendations to drivers in under a set of circumstances.

The Data

Data Procurement:

Data was received from the course website, through IBM. The dataset is assumed to be accurate and from a reputable source, and is confirmed from a check of the meta-data.

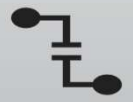
Data Information:

The dataset contains 37 total variables, 34 of which are independent.

3 of the variables are independent, but represent the same output.

- We address this as part of the data preparation.

There are 190,000+ rows, giving us more than enough data to run this analysis.



Data Issues:

Data is missing certain values.

Some of the data is represented in string format, slightly complicating the use of a decision tree to predict severity.

Handling of the Data

- The steps:
  1. Drop rows explicitly missing information.
     - 'EXCEPTRSNCODE' column, if it has 'NEI' as a value, means that there is data missing from that row.
  2. Convert simple strings ('Y'/'N') into numerical format ('0'/'1').
  3. Extract the relevant hour from the datetime column.
  4. Drop all columns that have been deemed to provide no impact on the outcome of the severity.
  5. Drop all rows that are missing data for any distinct value (not captured in step 1).
  6. Convert all remaining string based categorical variables into numerical values, and swap in place.
  7. Reset Index.

The Machine Learning Model

- Elected to use Decision Tree algorithm to predict severity.
- Steps:
    1. Separate total dataset into x and y variables (independent and dependent variables)
    2. Split the x/y variables into the trainsets and the testsets.
        - This will be split by 25% for testing, 75% for training.
    3. Run a for-loop to determine best output:
        - Create the Decision Tree object
        - Fit it to the training datasets
        - Predict the output for the remaining 25% of the x data (for testing purposes)
        - Calculate accuracy (as R^2).
            - If the accuracy has been improved:
                - Save as best accuracy
                - Record the depth of the tree for this step.

The Results

Calculate the three measures of accuracy:

Jaccard: ~.736	F1_Score: ~.715	R^2: ~.752

These scores are high enough for us to accept the model as accurate.

We can market this model to these business opportunities and create competitive advantage.

Conclusion

- Successful cleaning of the data to tease out valuable/impactful variables.
- Successful application of the data to a decision tree model.
- Successful test of the model to the existing testset.
  - Very high scoring on the accuracy
    - Important note: It is likely that this is not higher due to the large number of variables that contribute.

Further Exploration in the Future

- Locate dataset that includes additional complexity to the outputs:
  - Fatalities
  - Levels of Property Damage
  - Levels of Injury
- Manually adjust some of the data to allow for a more valuable correlation calculation.
  - Hour of day:
    - Instead of following the 24 hours of datetime starting at 12:00AM, adjust the 24 hours so that they start at 5:00AM. This would allow for 'increasing' numbers as we approach 5:00AM, so that 4:00AM would be 24 (the most dangerous hour).
    - Hypothetically, an accident at an abnormal time would have a higher severity, so adjusting for that might allow for a better correlation analysis.