A Model for Accident Forecasting by the Intersection

J.E.Turcotte | jturcott@gmu.edu
Department of Computational and Data Sciences | GMU Arlington

oscar.gmu.edu

RESEARCH QUESTION

Here, we ask if we can forecast future vehicular accident trends, by the intersection. We would like an application that could be used by municipalities when attempting to plan and manage civil and emergency response resources.

ABSTRACT

Utilizing a data set of roughly 3.3 million individual crashed vehicles across thousands of U.K. intersections over ten years from 2005 through 2014, a model and application has been developed that offers some potential practical insight. This involved significant complexity reduction through factor reduction, principal component analysis and general linear regression. The results were tested and refactored quite a few times and are a usable, if somewhat predictable, application.

ANALYSIS

Several iterations of data exploration occurred en route to the delivered model. Available in the data were capacities to assess details about the casualties and the vehicles involved in each recorded crash, but in addressing the nature of any given intersection, it was clear that much of this need not be included.

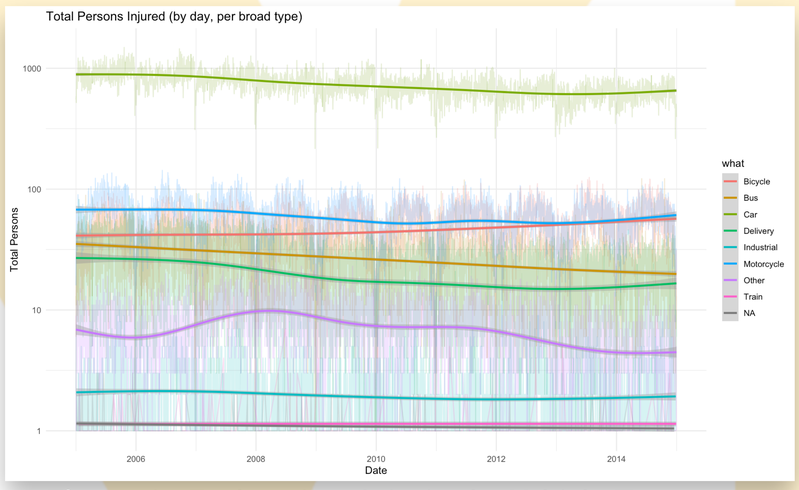

Figure 1 - Exploration: Injuries by Vehicle Type

Reducing factors down to the type of each road, the speed limit for each, the junction type (reduced in complexity down to stop vs. merge), proved to retain similar results to more complex tests while consuming fewer resources.

PROCESS

Principal component analysis, random forest, linear and generalized linear regressions were deployed in attempts to guess the severity of known accidents as model validation. 10-fold cross validations were used to train the final model which used 15 principal components and a generalized linear fit.

Examination of the results, seen in ***figure 2*** suggest that 'pretty good' predictions were possible a large majority of the time.

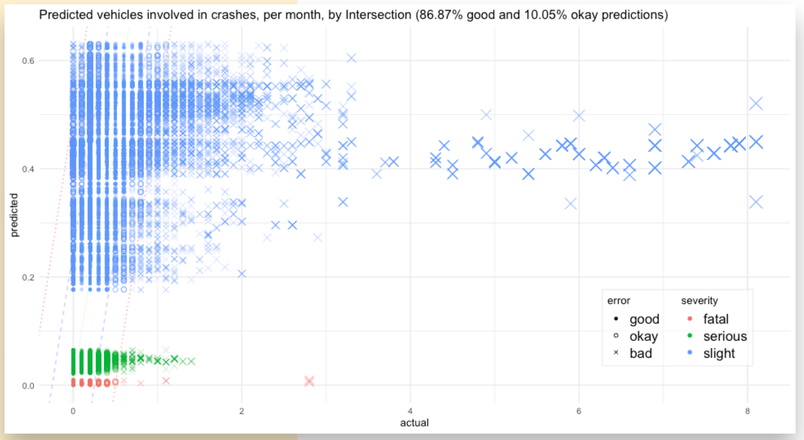

Figure 2 - Test Results with ~87% close predictions

Repeated testing and re-examination of the fits and their predictions revealed several errors made along the way, subsequently corrected, in terms of predictive factors that should have been used and were not, or were and should not.

The model itself accumulates its predicted accident tallies by examining the intersection counts coming from both directions... from the A-Class onto the B-Class and vice versa, since their scenarios will be different and have different trends.

FUTURE EFFORTS

One thing to note from ***figure 3*** (right), especially when testing multiple intersections, is that there is not much variation in the results. In fact, the strongest variation is from month to month, almost as if other conditions are not very relevant. It may be worth greater study to verify this.

CONCLUSION

A usable model has been produced that appears to correctly guess the number of accidents (per month) an appreciable percentage of the time, though the results appear to be rather variation-light.

Actual use of this model would be spread out across EVERY intersection in any given municipality, summed for desired grand totals, though it might also give insight into the nature of any one intersection.

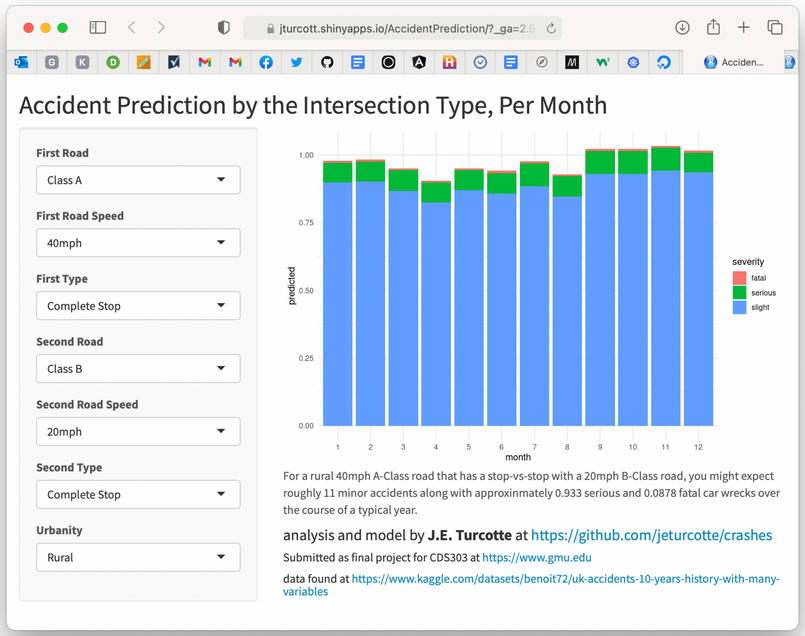

Figure 3 - Intersection Between A and B Class Roads

ACKNOWLEDGEMENTS

Much of this work would have been unapproachable without the education imparted by Dr. Carmen A. Iasiello, as assisted by Samiul Islam.

REFERENCES:

Luis Torgo. 2010. Data Mining with R: Learning with Case Studies (1st. ed.). Chapman & Hall/CRC.

Alfonso Zamora Saiz, Carlos Quesada González, Lluís Hurtado Gil, Diego Mondéjar Ruiz. 2020. An Introduction to Data Analysis in R. Springer Nature Switzerland AG 2020

Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) CRISP-DM 1.0.