MACHINE LEARNING FOR SAFER ROADS

Haridut Athi | Faculty Advisor: Prof Richard Sowers

Department of Industrial and Enterprise Systems Engineering, College of Engineering, University of Illinois at Urbana-Champaign

INTRODUCTION AND MOTIVATION

Vision Zero is a multi-national road safety project that aims to achieve zero road fatalities and drastically reduce serious injuries. New York and Los Angeles have been at the forefront in implementing this scheme and using data science to achieve the same. The traffic department of both the cities have made public instances of all the collisions over the recent years, which has inspired us to leverage the power of data to mine the patterns in fatal collisions.

The poster highlights some of the interesting trends obtained out of a preliminary analysis on Los Angeles collisions data and provides an insight as to how Logistic regression and Random Forest classifiers were used to predict fatal and serious collisions. It also throws light as to how we went about taking care of the class imbalance problem

MEET THE DATA

The Los Angles collisions data spans

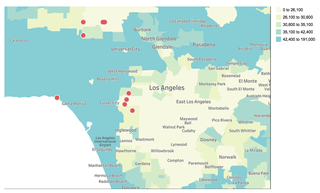
- ❑ 89 features
- ❑ 1.5 million collision points from 2009 to 2013

The features describe the circumstance of each collision and they talk about

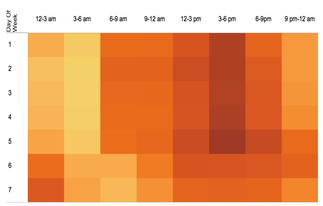
- ❑ Time, Day and Month of collision
- ❑ Geographical details surrounding a collision point
- ❑ Victim and Vehicle information

VISUALIZATIONS- TALKING POINTS

Social progress indices impacting road safety ?

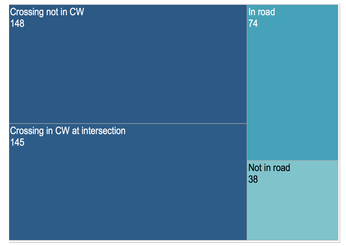


9 out of the top 10 collision spots in Los Angeles city are in the lower bracket when it comes to the per-capita income and even for other social indices

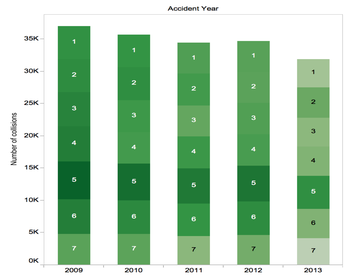


It turns out as one might expect that the collisions are more frequent between 3 to 6 pm during the week-days. There seems to be a spike in collisions during the weekends in the later hours.

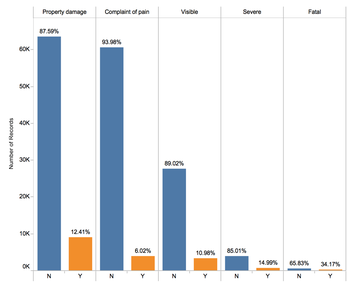
Over-speeding and automobile right of way dominate when it comes to the leading factors for fatal accidents.



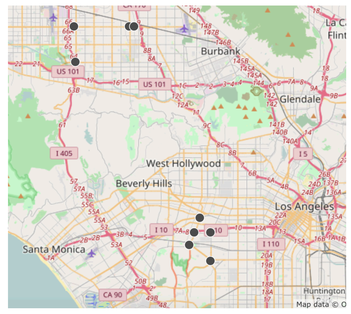
One interesting aspect that became apparent was left turn turned out to be one of the more dangerous manoeuvres when it comes to passenger – vehicle collision.



Friday consistently records the most collisions each year which is very interesting



The cases where the collisions are more fatal have more likelihood of a driver under the influence of alcohol being involved.



Three of the top 10 collision spots are on the lane connecting Van Nuyes Airport and Bob Hope Airport.

Unsafe speeds and Automobile right of way are the primary reasons for collisions at these spots too.

THE CLASS IMBALANCE PROBLEM

The fact that the class is heavily imbalanced towards the non-fatality cases it becomes very difficult for a machine learning model to predict the minority class of interest 'fatalities'.

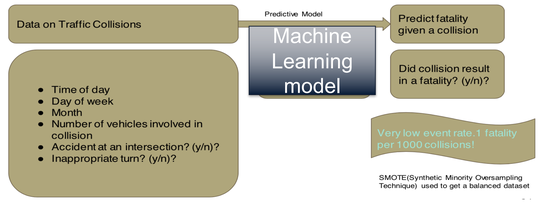
The Synthetic Minority Oversampling Technique (SMOTE) was used to generate synthetic samples to get over this problem.

In general, the algorithm selects two or more similar instances from the minor class (using a distance measure) and perturbs an instance one attribute at a time by a random amount within the difference to the neighboring instances.

When the data is categorical as in our case

- ❑ Over-samples the minority class and under-samples a majority class
- ❑ Select all the minority class instances, find the k-nearest neighbors among them
- ❑ For each of the feature, take the majority vote among the neighbors to generate a synthetic sample
- ❑ After performing SMOTE, the minority class was well represented with up to 40% of the data

PREDICTING FATAL COLLISIONS



Logistic regression model with LASSO regularization was used to select features that were appropriate and fit a linear classifier.

- ❑ The model was fitted with around with around 25 features
- ❑ The model was trained on the data generated by SMOTE
- ❑ The regularization parameter selected through a validation set was 0.1. The parameter was fine-tuned so that the model could achieve the highest recall
- ❑ The model was tested on a test set which was 400000 collision instances

Random Forest Classifier was also developed to find the most important factors in terms of the Gini index.

- ❑ The model used the same features as the logistic regression model
- ❑ Based on training the model on the validation set, 100 estimators or decision trees was chosen
- ❑ The depth of each decision tree was 9

INTERPRETATION AND RESULTS

- ❑ The odds of a fatal collision is 7 times more when the motorist crashes into a pedestrian who is on the sidewalk as opposed to a pedestrian on road
- ❑ The odds of a collision being fatal is 6 times more when alcohol is involved
- ❑ When it is dark and street lights are non-functional fatal collisions were more likely
- ❑ The top 3 factors based on the Gini Index obtained from the random forest model are Alcohol involvement, Rain and pedestrian involved in a collision with a motorist when he/she is not on the road but the side walk
- ❑ The Logistic regression model scored 84.6% on recall and 87% on precision
- ❑ The Random Forest model scored 92% on recall and 96% on precision

CURRENT WORK

- ❑ Given the network topology around the collision spot, we plan to predict the likelihood of a sever injury or fatality
- ❑ We then intend to identify the optimal network size that will minimize the error in predictions

ACKNOWLEDGEMENTS:

Thanks to the Los Angeles traffic authorities for making the data available and the data analyst team for their insights.

We would also like to thank my team-mates who worked on the New York City data and the under-graduate students who chipped in with their inputs.

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Template for a 48”x36” poster

Presenter name, Associates and Collaborators

Department of XXXXXXXXXXXXXXXX, College of XXXXXXXXXXXXXXXXX, University of Illinois at Urbana-Champaign

INTRODUCTION

This editable template is in the most common poster size (48” x 36”) and orientation (horizontal); check with the conference organizers for specific conference requirements regarding exact poster dimensions.

Writing Style:
The writing style for scientific posters should match the guidelines for your particular research discipline. Use the campus Writing Style Guide for general guidance with academic titles, names of campus buildings, the correct way to refer to the campus, etc.

Campus Guidelines
Authors should be aware of and follow the guidelines of the Institutional Review Board and the guidelines for campus copyright.

AIM

How to use this template
Highlight this text and replace it with new text from a Microsoft Word document or other text-editing program. The text size for body copy and headings and the typeface has been set for you. If you choose to change typefaces, use common ones such as Times, Arial, or Helvetica and keep the body text between 26 and 32 points.

The text boxes and photo boxes may be resized, eliminated, or added as necessary. The references to the department, college and university, including the logo, should remain.

Refer to the next page for logos commonly used on campus posters. You can drag and drop them to your personal PowerPoint scrapbook for use in subsequent posters; refer to PowerPoint help documents for more specific information regarding how to use the scrapbook.

METHOD

Text
Be sure to spell check all text and have trusted colleagues proofread the poster. In general, authors should:

- Use the active tense
- Simplify text by using bullet points
- Use colored graphs and charts
- Use bold to provide emphasis; avoid capitals and underlining
- Avoid long numerical tables

Authors should re-write their paper so that it is suitable for the brevity of the poster format. Respect your audience–as a general rule, less is more. Use a generous amount of white space to separate elements and avoid data overkill. Refer to Web sites or other sources to provide a more in-depth understanding of the research.

Captions set in a serif style font such as Times, 18 to 24 size, italic style.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat.

Captions set in a serif style font such as Times, 18 to 24 size, italic style.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat.

Captions set in a serif style font such as Times, 18 to 24 size, italic style.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat.

RESULTS

Images
TIFFs are the preferred file format for images appearing in printed posters. Avoid the use of low-resolution jpgs, especially those downloaded from the Internet, as they will reproduce poorly.

In order to insert an image, use the menu toolbar at the top of your screen.

Select:
1 Insert
2 Picture
3 From file
4 Find and select the correct file on your computer
5 Press OK

Be aware of the image size you are importing.

Captions set in a serif style font such as Times, 18 to 24 size, italic style.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat.

Captions set in a serif style font such as Times, 18 to 24 size, italic style.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat.

Captions set in a serif style font such as Times, 18 to 24 size, italic style.

Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat.

PRINTING & LAMINATING

Facilities & Services Printing Department will print and laminate posters in the dimensions of this template and provide a mailing tube for transportation at these prices:

$60.00 printing
$18.00 lamination
$5.00 proof (12.6" x 16.8”)
$3.50 mailing tube

To place your order, contact the Printing Department at 217-333-9350 or send an e-mail.

Please refer to estimate #005238 when submitting your order. Plan ahead; allow three business days for the Printing Department to complete the order. Other dimensions are available; the charge is by square foot. Contact the Printing Department for pricing information.

Resolving Printing Problems
PowerPoint does not always create the best PostScript files for printing. If you choose to have these printed on a campus plotter or by a third-party vendor and have printing errors, you may wish to export the file as a PDF and resend the file to the printing server/plotter.

CONCLUSIONS

We have created this template with scientific researchers in mind and with the help of feedback we have received. We encourage any comments or suggestions so that we can continue to update and improve this template. To make a suggestion contact:

creativeservices@illinois.edu

ACKNOWLEDGEMENTS

Check to make sure you’ ve acknowledged partner and funding agencies, either with text or with their logos.

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Commonly used logos