BikeSaferPA

Understanding severity of cyclist crash outcomes in PA, 2002-2021

Eamonn Tweedy, PhD

Data scientist candidate presentation
Department of Biomedical and Health Informatics
Children's Hospital of Philadelphia

Monday, August 14, 2023

Outline of the talk

1. Introduction to the problem and the PENNDOT dataset
2. Preparing data for modeling
3. Feature and model selection
4. Evaluating and explaining the model
5. Recommendations
6. Wrap up

Cycling for transit and recreation

Welcome to the
Benjamin Franklin Bridge
DELAWARE RIVER PORT AUTHORITY
RIGHT LANE ONLY
RADAR ENFORCED
PEDESTRIANS AND BICYCLES PROHIBITED ON ROADWAY

- Cycling is a sustainable and relatively low-cost way to get around and get more exercise
- Increasing the popularity of cycling can have positive public health and environmental impacts
  - *Policies that promote the use of bicycles should also address barriers such as fear of a traffic collision.*[1]

[1] WHO Cyclist Safety informational resource

Cyclist injuries and fatalities in the US

In the US annually:

- cyclists account for 2-3% of individuals killed and around 2% of individuals injured in motor vehicle crashes[2]
- cyclist fatalities have been above 800 since 2015, and neared 1000 in 2020 and 2021[1]
- the estimated total cost of bicycle injuries and deaths from crashes exceeds $23 billion[3]
- the burden on the healthcare system alone is over $400 million[4]

Goals of this project:

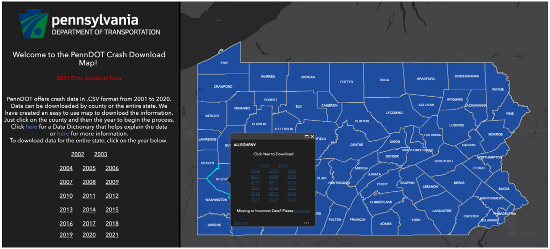
- **Build a classifier to predict whether a cyclist in a crash suffered serious injury or fatality**
- **Identify which features most heavily affect predictions, in order to target policy recommendations**

[2] NHTSA Fatality Analysis Reporting System (FARS)

[3] Centers for Disease Control and Prevention. Web-based Injury Statistics Query and Reporting System (WISQARS)

[4] Nationwide Inpatient Sample (NIS) database

The PENNDOT crash dataset



pennsylvania
DEPARTMENT OF TRANSPORTATION

Welcome to the PennDOT Crash Download Map!

2021 Data Available Now

PennDOT offers crash data in .CSV format from 2001 to 2020. Data can be downloaded by county or the entire state. We have created an easy to use map to download the information. Just click on the county and then the year to begin the process. Click here for a Data Dictionary that helps explain the data or here for more information.
To download data for the entire state, click on the year below.

2002 2003
2004 2005 2006
2007 2008 2009
2010 2011 2012
2013 2014 2015
2016 2017 2018
2019 2020 2021

ALLEGHENY

Click Year to Download

2002 2003
2004 2005 2006
2007 2008 2009
2010 2011 2012
2013 2014 2015
2016 2017 2018
2019 2020 2021

Missing or Incorrect Data? Please email us.

Zoom to

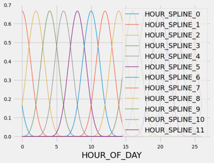
We use a publicly available crash dataset published by PENNDOT[5], covering all vehicle crashes in PA between 2002-2021; feature types:

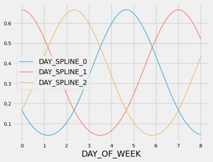
Crash-level	crash conditions, time, location, driver behavior, . . .
Vehicle-level	vehicle type, position/movement, role in crash, . . .
Person-level	age, sex, restraint/helmet, injury/fatality, . . .
Roadway-level	posted speed limit, lane count, . . .

[5] `https://crashinfo.penndot.gov/PCIT/welcome.html`

Data cleaning and feature engineering

- Several methods for imputing missing fields:
  - Using context from other features, e.g. weather ↔ road condition
  - Groupwise or dataset median/mode
    e.g. groupby (illumination,month) and use groupwise mode to fill missing hour-of-day
  - Creating “unknown” category in some cases
- Encoding features:
  - One-hot encoding for categorical features
    - n.b. - we can use Pandas category encoding for LightGBM models
  - Standard scaling for numerical features
  - Periodic basis spline encoding for cyclical features

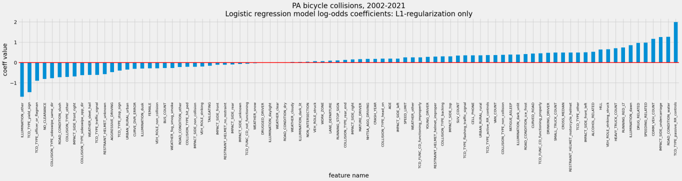




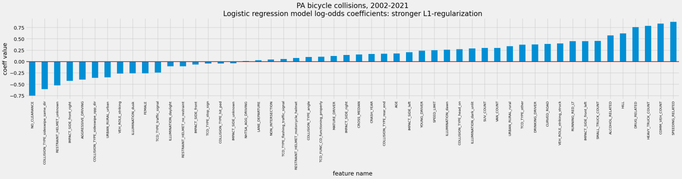
Feature selection via logistic regression coefficients

Model-assisted feature selection using L^1-regularized logistic regression

- promotes sparsity in the coefficients → consolidates feature influence



↓



- *Confirmed this choice via repeated trials (random stratified samples)*

Final list of features used

The following features remain after our elimination process:

Categorical:	**Binary:**	**Numerical:**
helmet status, striking/struck, urban/rural, illumination, crash type, impact side, road grade, cyclist sex	midblock, curve, signal, stop sign, flashing signal, alcohol-related, drug-related, mature driver, young driver, speeding, agg. driving, lane departure, median crossing, run red light, run stop sign, tailgating, proc. w/o clearance	cyclist age crash year speed limit
		Ordinal: counts of SUVs, small trucks, heavy trucks, vans, commercial vehicles
		B-splines: day of week, hour of day

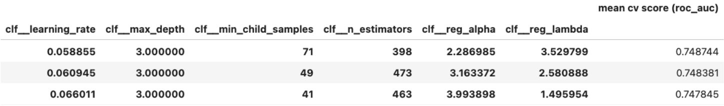
Selecting the BikeSaferPA model

Considered two model regimes:

- Logistic regression - tried L^1, L^2, and mixed regularization
- Tree-based models: standard GBDT model, and an optimized histogram-based leafwise-growth variant (LightGBM)

Selection process:

- Models evaluated using the ROC-AUC score
  - Area under the Receiver Operating Characteristic curve - plot of TPR vs. FPR as classification threshold is varied from $0 \to 1$
- Hyperparameter tuning: randomized search using repeated stratified cross-validation (5 folds, 3 repeats)
- **The winner: a tuned LightGBM model**

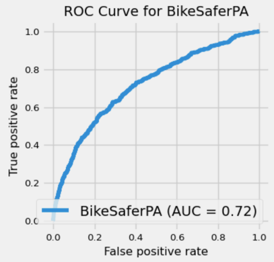


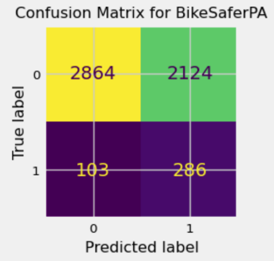
						mean cv score (roc_auc)
clf__learning_rate	clf__max_depth	clf__min_child_samples	clf__n_estimators	clf__reg_alpha	clf__reg_lambda	
0.058855	3.000000	71	398	2.286985	3.529799	0.748744
0.060945	3.000000	49	473	3.163372	2.580888	0.748381
0.066011	3.000000	41	463	3.993898	1.495954	0.747845

We selected the classification threshold that optimizes the F_3 score.

BikeSaferPA evaluation results

The holdout test set consists of 5377 samples (20% of the dataset)



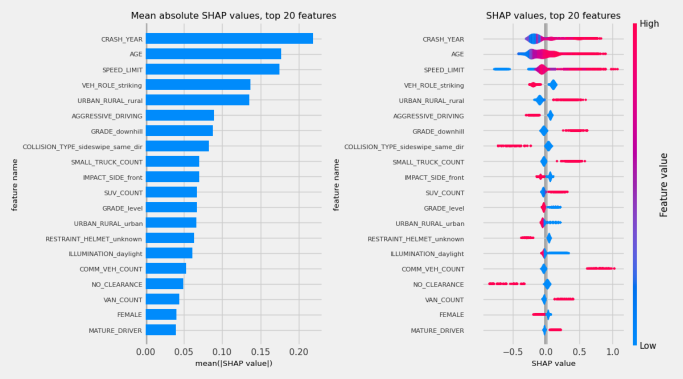


```
                                     precision    recall  f1-score   support

neither seriously injured nor killed       0.97      0.56      0.71      4988
         seriously injured or killed       0.12      0.75      0.20       389

                            accuracy                           0.58      5377
                           macro avg       0.54      0.66      0.46      5377
                        weighted avg       0.90      0.58      0.67      5377
```

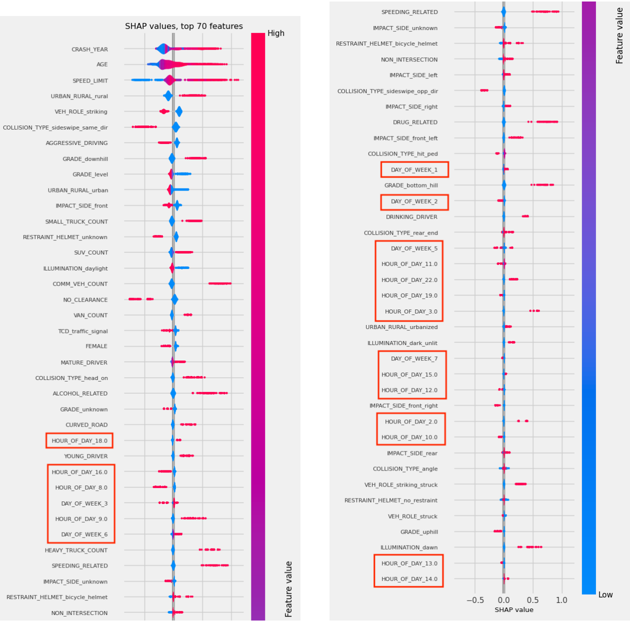
Explaining model predictions via SHAP values

SHapley Additive exPlanation (SHAP) values assign to each feature the average change in expected model prediction when adding that feature to the model

- From cooperative game theory
- Model agnostic!



Explaining model predictions via SHAP values



Policy recommendations

Recommendations based on my findings:

- Increase enforcement and motorist education related to speeding and impaired driving/riding
- Increase regulatory attention to large vehicles, e.g.
  - the number of such vehicles sharing the road with cyclists
  - speeds at which these vehicles may travel
  - required safety features and/or driver training
- Improve infrastructure, e.g.
  - upgrades/repairs to roadway lighting
  - adding protected bike lanes
- Increase motorists and cyclist education efforts regarding e.g.
  - safer biking practices around large vehicles
  - safer practices in higher speed zones and near hills, curves
  - use of lights and reflectors when riding in the dark
  - risks related to speeding and wrong-way riding
- Increase investigation of upward trend in severe injury and fatality
- *Strive to collect consistent, nuanced, and clean cyclist crash data nationwide*

Streamlit app demonstration

Try the app at either of the following:

`https://bike-safer-pa.streamlit.app/`

or

`https://huggingface.co/spaces/etweedy/BikeSaferPA`

Thanks for your attention!

Any questions?

References:

[1] Pennsylvania Department of Transportation. “Pennsylvania Crash Information Tool.” `https://crashinfo.penndot.gov/PCIT/welcome.html`

[2] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. *Scikit-learn: Machine Learning in Python. J Mach Learn Res.12. (2011), 329-2825–2830.*

[3] Ke, Guolin and Meng, Qi and Finley, Thomas and Wang, Taifeng and Chen, Wei and Ma, Weidong and Ye, Qiwei and Liu, Tie-Yan. *Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process.30 (2017), 3146–3154*.

[4] Lundberg, Scott M and Lee, Su-In. *A Unified Approach to Interpreting Model Predictions. Adv. Neural Inf. Process.30 (2017), 4765–4774*.

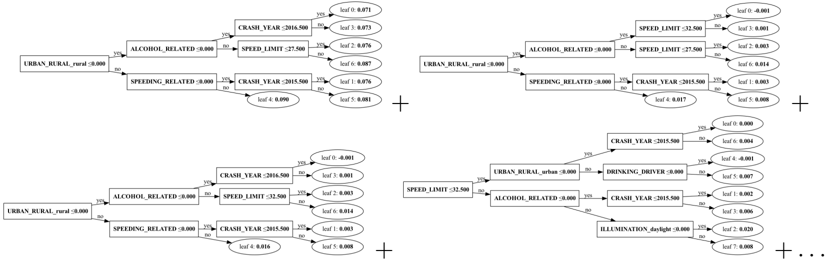
Appendix I: explaining model predictions via decision trees

Recall the gradient boosted decision tree training algorithm:

- A sequence of decision trees is trained, where each new tree is fit on the residual (error) of the prior stage prediction:

$$\begin{aligned} p_k &= p_{k-1} + \eta \cdot (tree_k.\mathrm{predict}(X)) \\ tree_{k+1} &= \mathrm{fit}(X, y - p_k) \end{aligned}$$

- Inference is done by summing the predictions of the entire sequence of trees



A handful of features reappear often in the beginning of the sequence and lead to relatively large contributions to the model's predictions

Appendix II: reflecting on model limitations

An AUC score of 0.72 is generally viewed as "acceptable" - would like to do better. There are some significant limitations in the dataset:

- Vehicle speed couldn't be used due to missing data
    - Speeding flag + posted speed limit was imperfect proxy
- A lot of details we don't know:
    - cyclist health condition
    - fine details of surroundings or impact event
    - severity/degree of binary features
- Significant "noise" in the target feature:
    - Many samples with "unknown" (24%) or "possible" (42%) injury status, which we did not consider as serious injury or death.
    - Samples with very similar input features but differing target values
        - 536 multi-cyclist collisions w/ mixture of outcomes
        - 373 cycles with passengers w/ mixture of outcomes