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

Outcomes for cyclists in crashes

- Cyclists account for 2-3% of individuals killed and around 2% of individuals injured in US motor vehicle crashes annually[1]
- Annual cyclist fatalities in the US has been above 800 since 2015, and neared 1000 in 2020 and 2021.[1]
- The total annual cost of bicycle injuries and deaths from crashes has been estimated to exceed $23 billion in the US.[2]
    - Burden to healthcare system along is over $400 million annually.[3]
- Increasing cycling as a mode of transit 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.[4]

It is crucial that policymakers, motorists, and cyclists better understand the factors that influence cyclist crash outcomes!

[1] NHTSA Fatality Analysis Reporting System (FARS)

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

[3] Nationwide Inpatient Sample (NIS) database

[4] WHO Cyclist Safety informational resource

The PENNDOT crash dataset

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

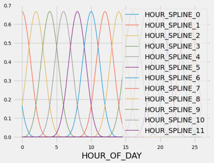
Crash-level	crash conditions, time, location, driver behavior, ...
Vehicle-level	vehicle type, position/movement, role in crash, ...
Person-level	age, sex, restraint/helmet, injury/fatality, ...
Roadway-level	posted speed limit, lane count, ...

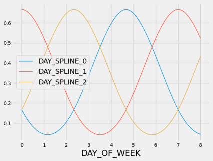
Our goals:

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

Data cleaning and feature engineering

- Several methods for imputing missing fields:
  - Using context from other features, e.g. weather $\leftrightarrow$ 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 use ordinal encoding with LightGBM model
  - Standard scaling for numerical features
  - Periodic basis spline encoding for cyclical features

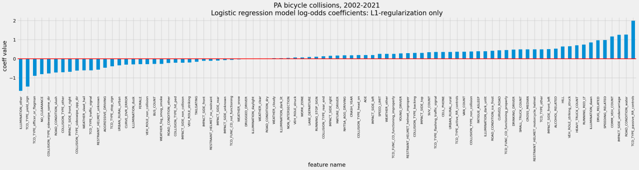




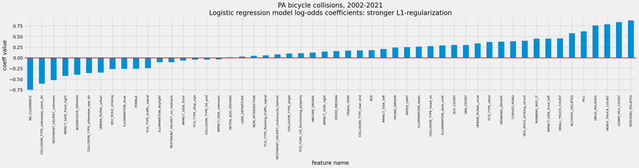
Feature selection via logistic regression coefficients

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

- promotes sparsity in the coefficients $\rightarrow$ consolidates feature influence



$\downarrow$



- *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 cyclist sex	midblock, curve, hill, 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
		Splines: day of week, hour of day

Selecting the BikeSaferPA model

Considered two model regimes:

- Logistic regression, with L^1, L^2, or mixed regularization
- Tree-based models: standard gradient boosted decision tree model, and an optimized histogram-based leafwise-growth variant (LightGBM)

Selection process:

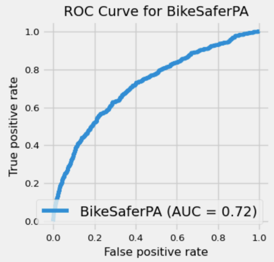
- Models evaluated using the ROC-AUC score
  - Area under the Receiver Operating Characteristic curve - plot of TPR vs. FPR as classification threshold is varied from $0 \to 1$
- Hyperparameter tuning: randomized search using repeated stratified cross-validation (5 folds, 3 repeats)
- **The winner: a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

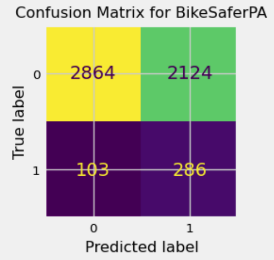
						mean cv score (roc_auc)
clf__learning_rate	clf__max_depth	clf__min_child_samples	clf__n_estimators	clf__reg_alpha	clf__reg_lambda	
0.058855	3.000000	71	398	2.286985	3.529799	0.748744
0.060945	3.000000	49	473	3.163372	2.580888	0.748381
0.066011	3.000000	41	463	3.993898	1.495954	0.747845

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

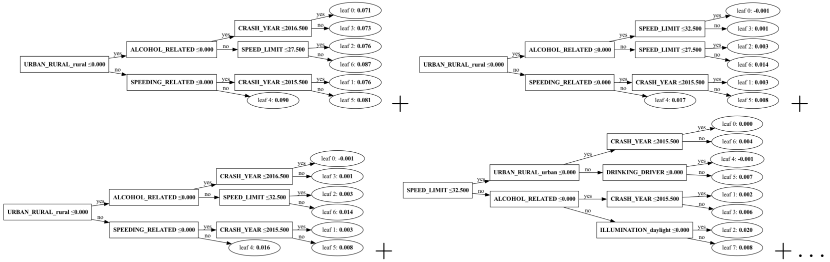
Explaining model predictions via decision trees

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:

$$p_k = p_{k-1} + \eta \cdot (tree_k.\text{predict}(X))$$
$$tree_{k+1} = \text{fit}(X, y - p_k)$$

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

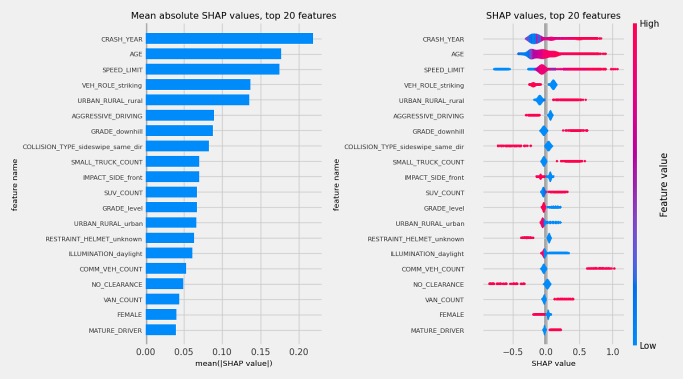


A handful of features reappear often in the beginning of the sequence and lead to relatively large contributions to the model's predictions

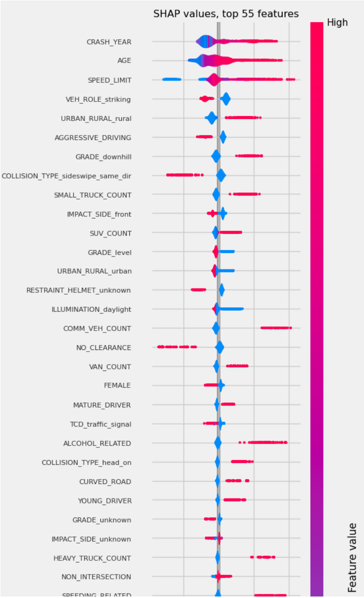
Explaining model predictions via SHAP values

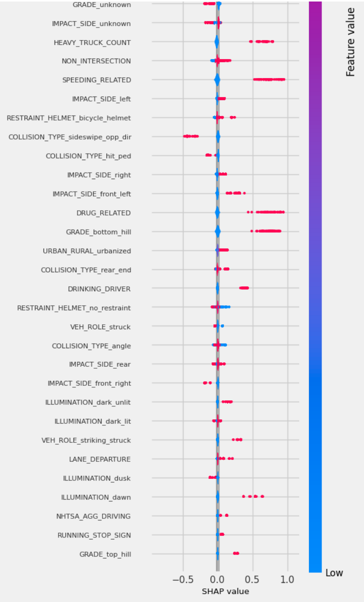
SHapley Additive exPlanation (SHAP) values assign to each feature the average change in expected model prediction when adding that feature to the model

- From cooperative game theory
- Model agnostic!

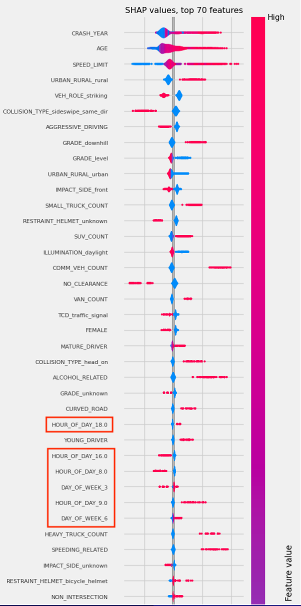


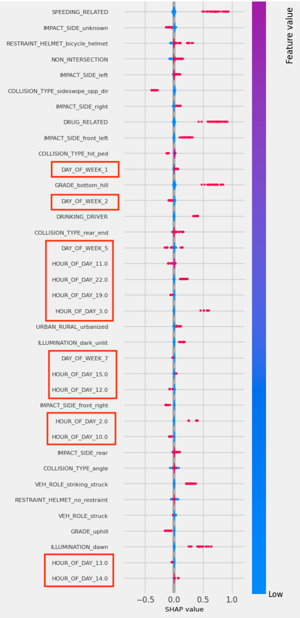
Explaining model predictions via SHAP values





SHAP values including day and hour features





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
- Many samples with "unknown" (24%) or "possible" (42%) injury status, which we did not consider as serious injury or death.
  - Adds "noise" to the target feature.
- No knowledge of cyclist health condition
- Behavior flags are non-specific
- No knowledge of fine details of surroundings, e.g. visibility, signage
- Many features can vary by degrees, some of which may influence outcomes significantly
- Multi-cylist crashes and multi-passenger cycles
  - 536 multi-cyclist collisions w/ mixture of outcomes
  - 373 cycles with passengers w/ mixture of outcomes

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

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*.