Motorcycle Collisions in Los Angeles

ALEX TRICKEY, PHD

3/14/2017

About Me

Education

- Dual BS, Math and Psych
- PhD, Quantitative Psychology

Current

- Data and Analytics Consultant
- Motorcycle Enthusiast

Image Source: https://www.pinterest.com/explore/new-motorcycles/

Motivation

City of Los Angeles - Vision Zero Initiative

- Goal: Eliminate deaths due to traffic incidents by 2025.
- Strategy: Use data to identify problems and guide decisions http://visionzero.lacity.org/

Objectives of Current Analysis

- **Identify leading predictors of severe motorcycle accidents and prototype a predictive model.**
- To the degree possible, establish data-backed recommendations for motorcyclists.

Analysis Overview

Acquire, transform, and explore the available data.

Establish a baseline model (GLM) to benchmark prediction quality and explore the data.

Tune and compare additional models (Random Forests and GBM).

Examine the results of modelling to advise motorcyclists.

Primary Data Source

Statewide Integrated Traffic Records System (SWITRS)

- Contains public records of collisions filed by the CHP and affiliated agencies.
- http://iswitrs.chp.ca.gov

Present Analysis Includes

- Records for collisions that involved a motorcycle in Los Angeles from Jan 2012 – Feb 2017.
- The resulting dataset contains 10,533 records.

Available Outcomes of Interest

Outcome

- The dataset does not contain non-collision data, so predicting accidents is not an option.
- Instead we can predict accident severity.
  - Severe Accidents: Involve a fatality or at least one severe injury
  - Non-severe Accidents: Involve only minor injuries or property damage

This outcome is imbalanced:

- Severe Accidents: 9,000
- Non-Severe Accidents: 1,533

Available Features of Interest

Examples of Available Features

- When and where: Date/time of the accident, Did the accident occur at an intersection? On a highway?
- Accident details: Traffic violations, kinds of vehicles involved, type of collision
- Conditions: Weather, road conditions, lighting

Baseline/Exploratory Model

A GLM (Elastic-Net) Model was used to examine predictive potential and the relationships between the features and the outcome.

Model Quality Metric

- The AUC was used as the primary metric of model prediction quality.
  - Represents the probability that the model will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example.
  - Relatively robust to class imbalance.

Fitting the Elastic-Net GLM

Elastic-Net Regression is similar to standard regression, but uses a regularization term to reduce the weight of weak predictors and reduce overfitting.

All models were implemented in using the h2o R library.

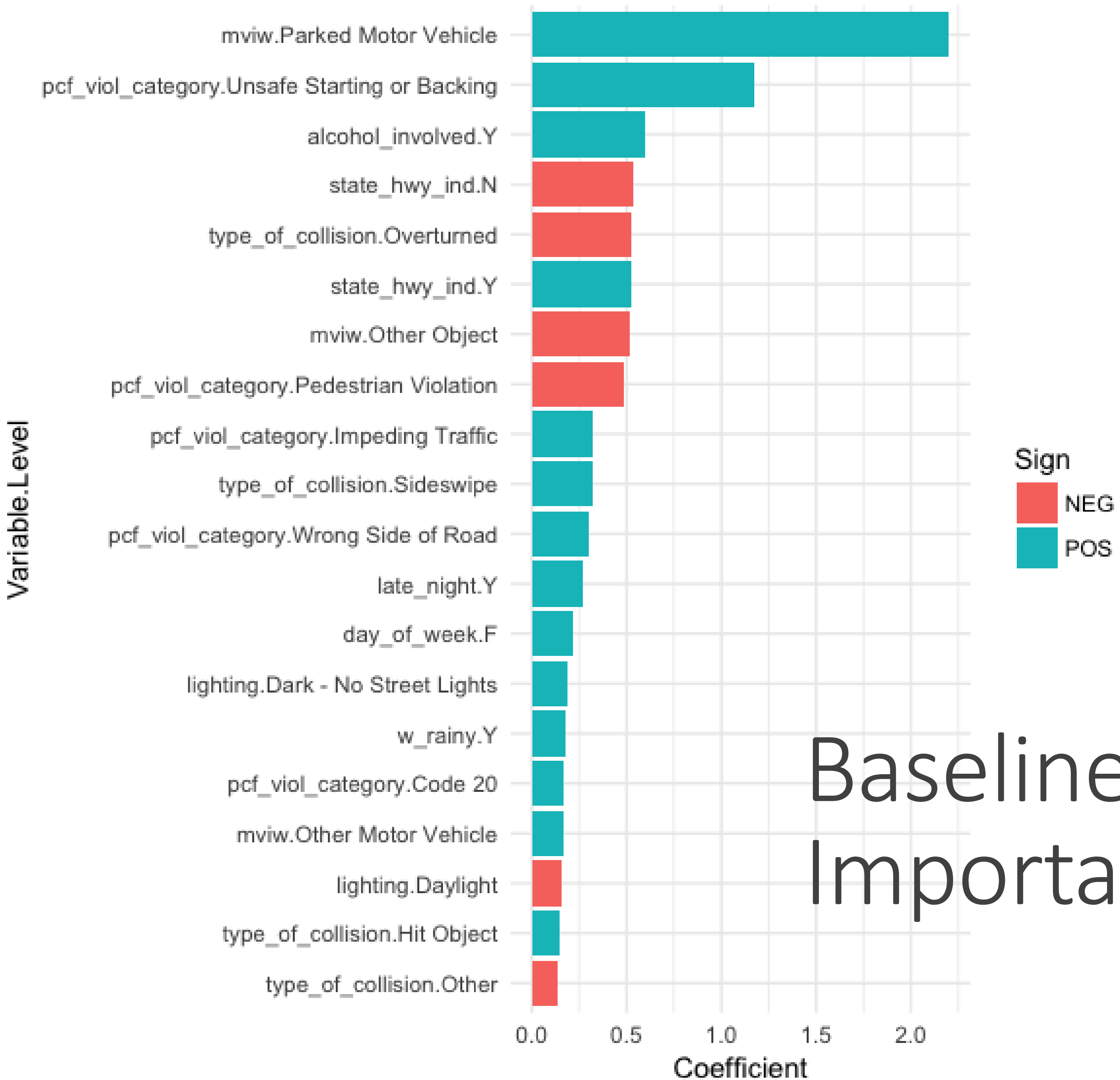
A 5-fold cross validation scheme was used to tune the λ regularization parameter using the training data.

Baseline Prediction Quality

Prediction Quality

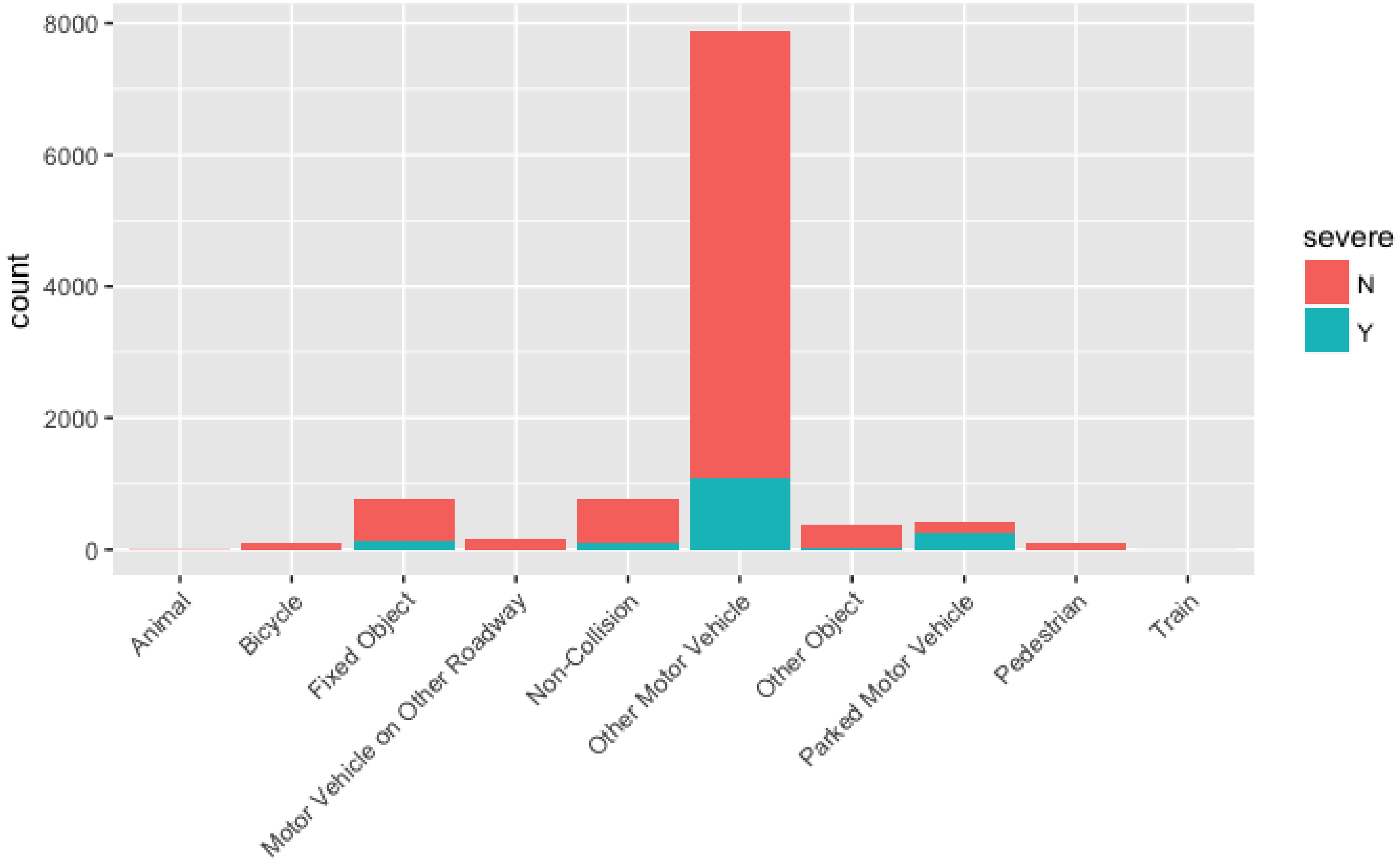
- AUC: 0.7259
- Mean Per-Class Error: 0.338
- Validation Confusion Matrix (uses max F1 threshold)

	Predicted Not Severe	**Predicted Severe**	**Error-Rate**
Actually Not Severe	1113	204	0.155
Actually Severe	126	116	0.521
Totals	1239	320	0.212



Baseline Feature Importances

Follow-up on “Motor Vehicle Involved With” (MVIW)



Gradient Boosted Machines

The Gradient (Tree) Boosting Machine is an ensemble method which builds a series of models that progressively improve upon previous models.

Each model is effectively fit to the residual error of previous models, to additively built a superior model.

References:

- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
- H2O Implementation: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/gbm.html#gbm-algorithm

Fitting the GBM

A validation set and a random discrete search procedure were used to tune the model hyperparameters using h2o.grid().

Some Parameters Explored:

- The learning rate
- The max depth of the trees
- The number of trees
- Class sampling factors

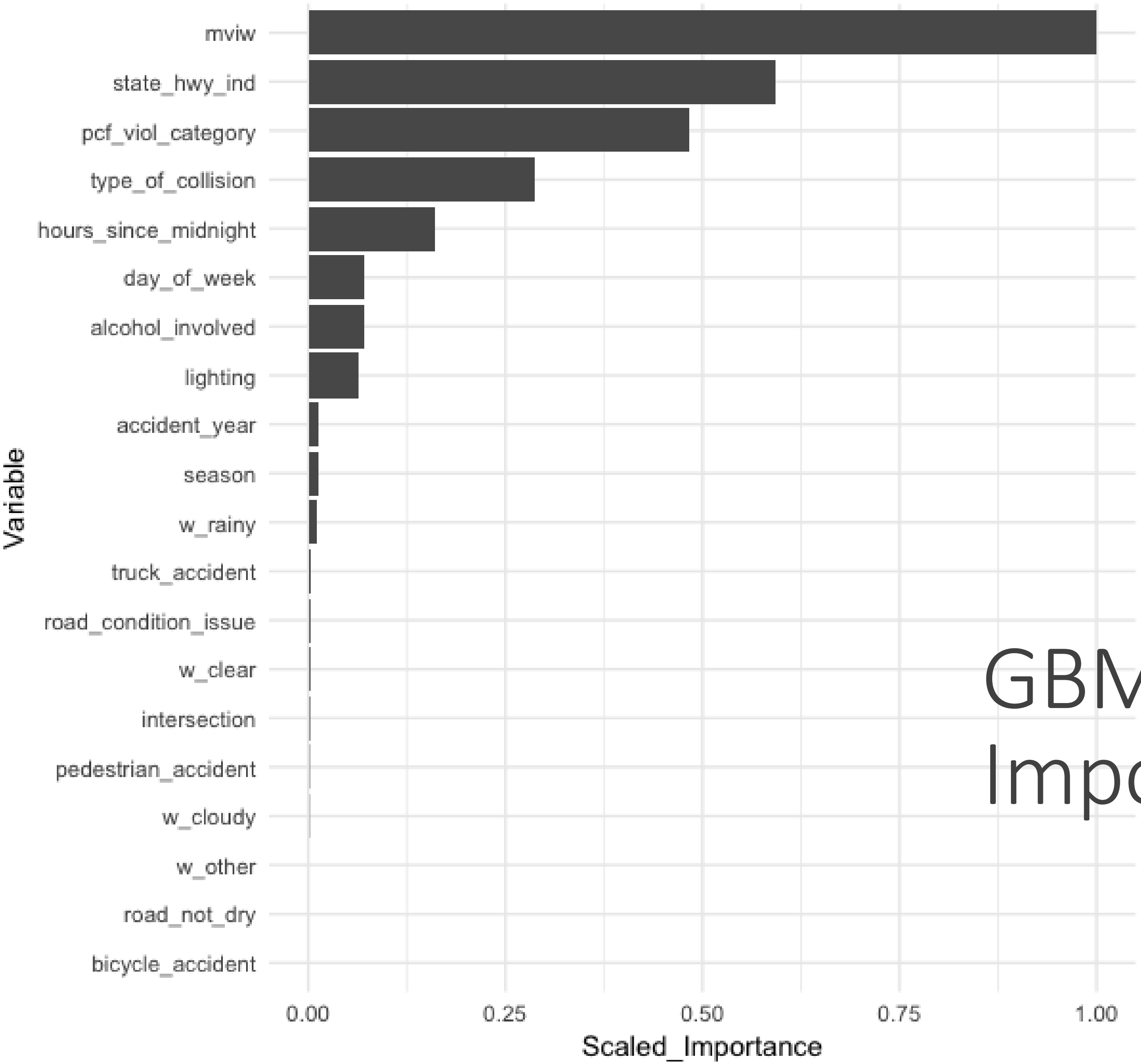
The model which returned the largest AUC was selected for consideration.

GBM Prediction Quality

Final results in the hold out set:

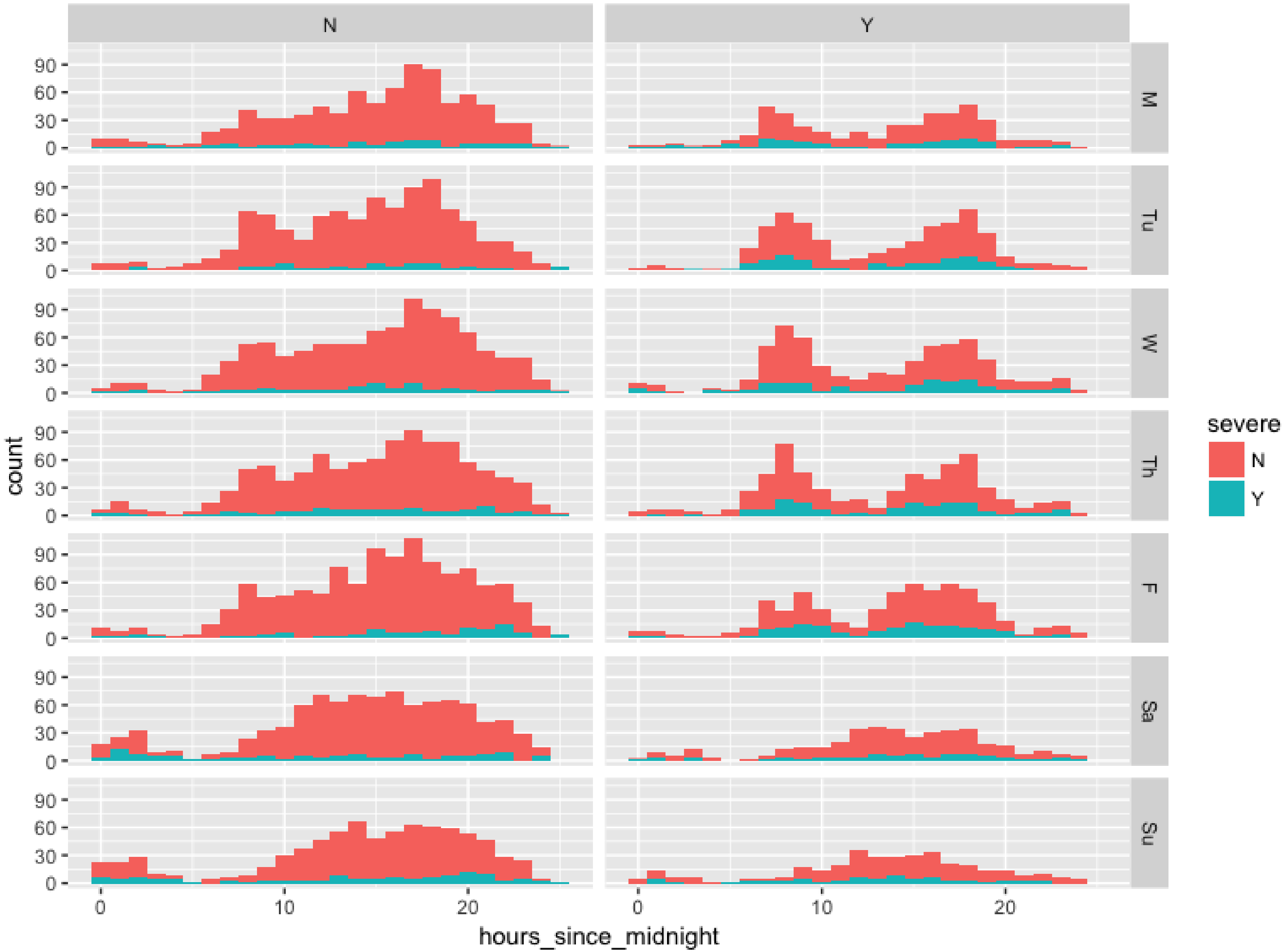
- AUC: 0.7566
- Mean Per-Class Error: 0.316
- Confusion Matrix (uses max F1 threshold)

	Predicted Not Severe	**Predicted Severe**	**Error-Rate**
Actually Not Severe	1089	269	0.198
Actually Severe	88	115	0.433
Totals	1177	384	0.228



GBM Feature Importances

Highway or Local Streets by Day/Time



Limitations and Next Steps

Unreported Data

- Missing equivalent data for motorcycles that are not in accidents and data for accidents without a police report.
- Estimates of traffic volume, motorcycle usage, frequency of unreported accidents would help.

Next Steps

- Continue to clean and supplement the dataset.
- Incorporate specific location data, road quality data, traffic volume data.
- Consider alternative outcomes (e.g. number of fatalities, Likhert scale severity rating, Number Injured)

Conclusions

It is possible to build predictive models of motorcycle accident severity.

Further work is needed to ensure that these models have sufficient data to yield high-quality predictions.

Regardless, the exploratory and modeling analyses presented here provide much needed insight into

- Which motorcycle accidents are most dangerous
- When and where bikers should drive (or not drive) to minimize risk of severe injury

Thank You!

QUESTIONS?