Prospectus Report: Building a Model to Recommend Dispatching an Ambulance based on Automated Crash Reports from Cell Phones

Brad Burkman

Updated 13 October 2022

One-Page Summary

Problem: Given an automated crash report from a cell phone, use two historical datasets to build and analyze models to recommend whether to immediately dispatch an ambulance. The solution requires data cleaning, imputing unknown values, and handling imbalanced data.

1. I am qualified and on track to graduate in December 2023. See Qualifications.
2. My dissertation will demonstrate competence in the major techniques of research.
   a. Finding an interesting question whose answer requires current and novel methods
   b. Literature review
   c. Finding appropriate datasets
   d. Data cleaning and imputation of missing values
   e. Handling imbalanced data
   f. Building, testing, and comparing models
   g. Analysis of results in terms of the application
   h. Analysis of results in terms of the current and novel methods
3. My dissertation will demonstrate that I have wrestled with the data in these aspects that are more art than science.
   a. Imputing missing values
   b. Binning (discretizing, batching) many categories into fewer
   c. Handling imbalanced data
4. My dissertation will make novel contributions to the field.
   a. Novel application
   b. Previously unused method for imputing unknown values in a major dataset
   c. New metrics: Balanced precision and balanced F1
   d. New interpretation of class weights as a political/ethical cost/benefit tradeoff
5. I have reviewed the literature in these areas.
   a. ML metrics for imbalanced data
   b. Dataset balancing techniques
   c. Others' use of the CRSS dataset and how they handled missing data
   d. Use of the metrics and imbalanced data techniques in the crash analysis literature
6. I am preparing a paper for submission to a respected journal. See accompanying draft and the rankings of journals.
7. I have a detailed and realistic plan for completing the dissertation. See Research Plan.
8. This document and the accompanying paper draft illustrate that I can make useful large documents to different specifications.
9. I have more questions than answers.

Contents

Index

Acronyms

CRSS Crash Report Sampling System. 14, 15, 58, 63, 66, 86

DOT Department of Transportation. 14, 58

IVEware Imputation and Variance Estimation Software. 68, 86

NASS GES National Automotive Sampling System General Estimates System. 58
NHTSB National Highway Transportation Safety Board. 14, 58, 66

SHRP2 Second Highway Research Program. 50
SMOTE Synthetic Minority Oversampling Technique. 43
SMRI Sequential Regression Multivariate Imputation. 69

Chapter 0

Qualifications

0.1 Preparation and Degree Plan

0.1.1 Previous Education

1989 - 1993	Wheaton College (IL)
	B.A. in English and Economics (double major)
1998 - 2000	SUNY Buffalo
	M.A. Mathematics
	Returned 2001-03 for additional coursework (total 60 hours)
	Passed first PhD Qualifying Exam

0.1.2 Courses Taken

Transfer - UBuffalo	3	MATH 595	PDE's
Transfer - UBuffalo	3	MATH 555	Numerical Analysis I
Transfer - UBuffalo	3	MATH 556	Numerical Analysis II
Transfer - LSU Shreveport	3	CSCE 502	Bioinformatics
Fall 2018	3	CSCE 515	Graphics
Fall 2018	3	CSCE 553	Software Methodology
Fall 2018	3	CSCE 561	Information Storage and Retrieval
Fall 2018	1	CSCE 595	Seminar
Fall 2018	3	CSCE 669	Raghavan Adv. Topics
Spring 2019	3	CSCE 500	Algorithms
Spring 2019	3	CSCE 509	Pattern Recognition
Spring 2019	3	CSCE 530	Architecture
Spring 2019	1	CSCE 595	Seminar
Fall 2019	3	CSCE 572	Combinatorial and Geometric Algorithms
Fall 2019	1	CSCE 595	Seminar
Spring 2020	3	CSCE 619	Jin Adv. Topics
Spring 2020	1	CSCE 595	Seminar
Fall 2020	3	CSCE 619	Jin Adv. Topics
Fall 2020	1	CSCE 595	Seminar
Spring 2021	3	CSCE 619	Jin Adv. Topics
Summer 2021	3	CSCE 619	Jin Adv. Topics
Fall 2021	3	CSCE 699	Jin Dissertation
Spring 2022	3	CSCE 699	Jin Dissertation
Summer 2022	3	CSCE 699	Jin Dissertation
Total (excluding 595)	57		
Total 595	5		

0.1.3 Examinations

- GRE (19 May 2016)
  - 170 Quantitative
  - 170 Verbal
- PhD Comprehensive Exams
  - Software Engineering (January 2019)
  - Algorithms (August 2019)

0.1.4 PhD Degree Requirements

	√	CSCE 500
	√	Breadth Requirement
		One 500-level course in hardware
		CSCE 530
		Two 500-level courses in software
		CSCE 553 Software Methodology
		CSCE 561 Information Storage and Retrieval
		One 500-level course in theory
		CSCE 500 Algorithms
		One other 500-level course in areas not listed above
		CSCE 515 Graphics
		Any accepted 500-level course
		CSCE 509 Pattern Recognition
7, 3.85	√	Six 500-level courses in CACS with a GPA of at least 3.5
12	√	At least 9 hours of CSCE 6x9 research courses
	√	PhD Comprehensive Exam
		Software Engineering (January 2019)
		Algorithms (August 2019)
	×	PhD Prospectus Exam
	×	PhD Dissertation Defense
9	×	Exactly 24 hours of CSCE 699 (dissertation credit)
48	√	48 other hours
5	√	5 semesters of CSCE 595

0.1.5 Fall 2022 Plan

Fall 2022	3	CSCE 699	Dissertation
		Prospectus Exam	(tentatively Friday 4 November 2022)

0.1.6 Remaining Requirements

12 hours of 699
PhD Dissertation Defense

0.1.7 Plan for Completing Degree

Six years from Fall 1998

Spring 2023	3 hours 699
Summer 2023	6 hours 699
Fall 2023	3 hours 699
	PhD Dissertation Defense (December 2023)

0.1.8 Committee Members

Dr. Henry Chu	CACS	Chair
Dr. Xiaoduan Sun	Civil Engineering	
Dr. Aminul Islam	CACS	
Dr. Mehmet Tozal	CACS	

0.2 Previous Work

Two documents accompany this prospectus report to show the variety of work I have done.

1. A partial draft of the paper I plan to submit to *Transportation Research Part C: Emerging Technologies* in January 2023, using the journal's LaTeXtemplate.
2. My study guide for the 2019 Algorithms qualifying exam.
   `https://github.com/bburkman/Algorithms_Comp_Prep/blob/2fabe0e05bb13118a58a83e55016f4158de19c9c/CSCE_500_Comps_Prep/Algorithms_Comp.pdf`

Chapter 1

Introduction

1.1 Problem

1.1.1 Application

New (starting in 2022) Google Pixel phones have a feature that will automatically alert the police when involved in an automobile crash. Apple says the feature is coming to iPhones and Apple Watches soon; those products already have a feature that detects a person falling, calls the person, and if no response, calls a neighbor, a friend, or the police. One of my friends with multiple sclerosis uses this app.

Such systems (like GM OnStar) , built into vehicles, have existed for years, but soon they will become ubiquitous. When the police receive a notification, based on the information they have, should they automatically deploy an ambulance? In an accident with severe (but not instantly fatal) injuries, a few minutes' delay may have serious consequences, but sending an ambulance is expensive, and their supply is limited. Can we develop a model that will, from the limited information the police can hope to have, from the datasets we have chosen, build a model to make a good prediction of whether an ambulance is needed?

I am using "police" as a shorthand for "the decision makers at the emergency call center."

This new cell phone feature will not be perfect; it will give many false positives and may not detect crashes with small objects, like pedestrians, that do not cause severe deceleration but are most likely to have severe injury. The automated reports may, however, give us additional information like the number of people (number of phones) involved, and speed at time of impact. This new phone feature will keep the crash analysis community busy for many years.

The "make a good predicition" part is complicated. We are not going to get 100% accuracy. What would we mean by "good," and what would we use as a basis of comparison? The current system relies mostly on phone calls from eyewitnesses who can give more information than the police will have in an automated notification. These are thorny questions that we must address.

1.1.2 Datasets

I am looking at two datasets, the US Department of Transportation (DOT) National Highway Transportation Safety Board (NHTSB) Crash Report Sampling System (CRSS) data 2016-2020 data ($\approx 250,000$ records), and a census of Louisiana crash records 2014-18 ($\approx 800,000$ records).

1.1.3 Imbalanced Data

In the 2014-2018 Louisiana data, we have over eight hundred thousand crash records. If we are just looking for fatal crashes, about 3500 were fatal, 0.42%. If we built a model to predict whether a crash is fatal, and the model predicted that all crashes were nonfatal, that model would have correctly classified 99.58% of crashes, or have 99.58% *accuracy.* In most contexts, that level of accuracy would be amazing, but in this context, such a model would be useless.

In the CRSS dataset, which over represents severe crashes, 81.15% of people involved in a crash were not transported to the hospital, and 16.75% went to the hospital (the remaining 2.10%

unknown). This nearly 5:1 imbalance is not as severe as the example with fatalities above, but still will be a challenge for our usual model building algorithms to give us the insights we want.

The problem of imbalanced data appears in many applications, including spam detection and credit card fraud detection, and over the last fifty years the community has built many tools for addressing the problem. Using those tools is as much art as science, and the best combination of methods depends on the dataset and desired outcome. The desired outcome is a moral, ethical, and political question as well as a technical one.

1.1.4 Tradeoffs

Balancing false positives and false negatives in this application is additionally problematic because they have different costs. The cost of a false positive (sending an ambulance when one is not needed) is measured in dollars, but the cost of a false negative (not sending an ambulance when one is needed) is measured in lives. It is likely that this study will only illustrate the choices to be made rather than find a Goldilocks solution that will significantly increase the number of true positives without increasing the number of false positives.

1.2 Novel Contributions of this Work (Knowledge Gap)

Novel Aspects of this Work

- New Real-World Problem: Newly emerging problem of how to use the greatly increasing volume of automated crash notification data.
- New Dataset: The Louisiana dataset has not appeared significantly in the literature.
- New Imputation of Unknown Values in Well Known Dataset (CRSS)
- New Metrics: Balanced Precision and Balanced F1
- Interpretation of Class Weights as a Political/Ethical Cost-Benefit Tradeoff
- New Combinations of Methods: The Louisiana data is very incomplete, dirty, and imbalanced, and the CRSS data is imbalanced. Off-the-shelf methods will not give the level of confidence needed for life-and-death decisions.

Chapter 2

Lit Review: Crash Analysis

2.1 Journals with Self Description and Rankings

Journal	CiteScore	Impact Factor
Accident Analysis and Prevention	7.8	4.993

Accident Analysis & Prevention provides wide coverage of the general areas relating to accidental injury and damage, including the pre-injury and immediate post-injury phases. Published papers deal with medical, legal, economic, educational, behavioral, theoretical or empirical aspects of transportation accidents, as well as with accidents at other sites. Selected topics within the scope of the Journal may include: studies of human, environmental and vehicular factors influencing the occurrence, type and severity of accidents and injury; the design, implementation and evaluation of countermeasures; biomechanics of impact and human tolerance limits to injury; modelling and statistical analysis of accident data; policy, planning and decision-making in safety.

American Journal of Emergency Medicine	3.2	2.469

A distinctive blend of practicality and scholarliness makes the American Journal of Emergency Medicine a key source for information on emergency medical care. Covering all activities concerned with emergency medicine, it is the journal to turn to for information to help increase the ability to understand, recognize and treat emergency conditions. Issues contain clinical articles, case reports, review articles, editorials, international notes, book reviews and more. The American Journal of Emergency Medicine is recommended for initial purchase in the Brandon-Hill study, Selected List of Books and Journals for the Small Medical Library (2001 Edition).

Decision Support Systems	10.5	5.795

The common thread of articles published in Decision Support Systems is their relevance to theoretical and technical issues in the support of enhanced decision making. The areas addressed may include foundations, functionality, interfaces, implementation, impacts, and evaluation of decision support systems (DSSs). Manuscripts may draw from diverse methods and methodologies, including those from decision theory, economics, econometrics, statistics, computer supported cooperative work, data base management, linguistics, management science, mathematical modeling, operations management, cognitive science, psychology, user interface management, and others. However, a manuscript focused on direct contributions to any of these related areas should be submitted to an outlet appropriate to the specific area.

Examples of research topics that would be appropriate for Decision Support Systems include the following:

1. DSS Foundations e.g. principles, concepts, and theories of enhanced decision making; formal languages and research methods enabling improvements in decision making. It is important that theory validation be carefully addressed.

2. DSS Functionality e.g. methods, tools, and techniques for developing thefunctional aspects of enhanced decision making; solver, model, and/or data management in DSSs; rule formulation and management in DSSs; DSS development and use in computer supported cooperative work, negotiation, research and product.

3. DSS Interfaces e.g. methods, tools, and techniques for designing and developing DSS interfaces; development, management, and presentation of knowledge in a DSS; coordination of a DSS's interface with its functionality.

4. DSS Implementation - experiences in DSS development and utilization; DSS management and updating; DSS instruction/training. A critical consideration must be how specific experiences provide more general implications.

5. DSS Evaluation and Impact e.g. evaluation metrics and processes; DSS impact on decision makers, organizational processes and performance.

Journal of Safety Research 5.0 3.487

The Journal of Safety Research is a multidisciplinary publication that provides for the exchange of scientific evidence in all areas of safety and health, including traffic, workplace, home, and community. While this research forum invites submissions using rigorous methodologies in all related areas, it focuses on basic and applied research in unintentional injury and illness prevention. Affiliated with the National Safety Council, it seeks to engage the global scientific community including academic researchers, engineers, government agencies, policy makers, corporate decision makers, safety professionals and practitioners, psychologists, social scientists, and public health professionals.

Transportation Research Part C: Emerging Technologies 14.0 8.089

The focus of Transportation Research: Part C (TR_C) is high-quality, scholarly research that addresses development, applications, and implications, in the field of transportation systems and emerging technologies . The interest is not in the individual technologies per se, but in their ultimate implications for the planning, design, operation, control, maintenance and rehabilitation of transportation systems, services and components. In other words, the intellectual core of the journal is on the transportation side, not on the technology side. The integration of quantitative methods from fields such as operations research, control systems, complex networks, computer science, artificial intelligence are encouraged.

Of particular interest are the impacts of emerging technologies on transportation system performance, in terms of monitoring, efficiency, safety, reliability, resource consumption and the environment. Submissions in the following areas of transportation are welcome: multimodal and intermodal transportation; on-demand transport; intelligent transportation systems; traffic and demand management; real-time operations;

connected and autonomous vehicles; logistics; railways; resource and infrastructure management; aviation; pedestrians and soft modes.

Special emphasis is given in open science initiatives and promoting the opening of large-scale datasets for papers published in TR_C that can support transferability and benchmarking of different approaches. The realization of data opportunities that arise from emerging technologies and new sensors in transportation can revolutionize how this data reshape our understanding of congestion mechanisms and can contribute in efficient and sustainable mobility management.

2.2 Articles using Similar Datasets

- Rahim 2021 [124] LSU faculty, similar dataset to what we have.
- Jiang 2020 [66] used similar data and addressed the challenges we'll have with it.

2.3 Articles on Imbalanced Crash Data

- Schlogl 2020 [131] uses imbalanced data.

2.4 Ambulances

From 11/29/21 Report.

- Found standard for emergency medical service (EMS) response time, from The National Fire Protection Association. "1710 NFPA Standard for the Organization and Deployment of Fire Suppression Operations, Emergency Medical Operations, and Special Operations to the Public by Career Fire Departments, 2020" §4.1.2.1
  - 60-second turnout time
  - 240 seconds or less travel time for the arrival of a unit with first responder with automatic external defibrillator (AED) or higher-level capability at an emergency medical incident
  - 480 seconds or less travel time for the arrival of an advanced life support (ALS) unit at an emergency medical incident, where this service is provided by the fire department provided a first responder with an AED or basic life support (BLS) unit arrived in 240 seconds or less travel time.
  - Lots of papers, like Liu (2016)[91] cite Rafael Sa'adah (2004), which I think is a response to the NFPA standards, but I can't find it online or in the library database.

2.5 iPhone to Automatically Detect Crash and Call Emergency Services

From 11/29/21 Report.

- iPhones and Apple Watches will soon automatically call police when the accelerometer detects a car crash.
- Several articles dated 11/1/21, including in the Wall Street Journal.
- Available in 2022
- What data would that provide, and what data would the police already have to complement it? These are just my guesses.
- Data from Apple
  - Registered owner of the phone (or phones) in the car
  - Typical users of that phone (Apple knows!)
  - GPS location
  - Perhaps a rough idea of how fast the car was going and how suddenly it stopped
  - If more than one phone sends signal, do these people know each other, or are they likely in different vehicles?
  - Accelerometer signature of a pedestrian or bicyclist getting hit?
- Complementary data from police database
  - Type of roadway and speed limit
  - Was it at an intersection?
  - Time of day, day of week
  - Type of vehicle registered to that person
  - Driving record of user of phone (History of DUI?)
  - Weather

2.6 Weather

From 11/29/21 Report.

- Wang et al [156] studied the data of a ride-hailing company, DiDi Chuxing, and looked for how resilient the system was during “extreme weather events.”
  - They defined such weather to be “hurricane, flooding, and rainstorm.” (page 2) I suspect that “rainstorm,” which is really vague in English, is a poor translation of a more specific Chinese word.
  - Because these extreme events are rare, they have a sample imbalance problem (page 13). They solve the problem in an interesting way, by ignoring it and watering down their data set. “The characteristics of urban transportation resilience under catastrophic events have generally similar patterns to those under general precipitation events. Thus we incorporated the rainstorm and usual prediction events data into [sic] data set to strengthen the model training.” So, as I understand it, they had an imbalanced data problem modeling extreme weather, so they just modeled ordinary weather.
  - I like how the authors started their methodology section with a page of definitions.

2.7 Lagniappe

- Osman 2019 (LSU) [112] looked much more deeply at the data than other studies, looking for correlations between sets of variables.
- Ziakopoulos 2020 [205] is a good overview of the field and its jargon.
- Guimmarra 2020 [52] is interesting for its text mining of crash reports.
- Park 2019 [117] has a full-page table categorizing studies of ambulance location, relocation, and dispatching using different optimization methods.

2.8 Significant Authors

From 11/15/21 Notes:

Reviewed all of the 66 articles from 2021 with the word “crash” in *Transportation Research Part C: Emerging Technologies.*

- Most of the articles are about autonomous vehicles.
- Mohammed Abdel-Aty at the U of Central Florida is a major author in this journal, but not in this year. In previous years, if there was an article from UCF, his name was on it. His website does not say that he has retired.
- When I write, I want to include more examples than many authors give.

2.9 TR_C Articles on Machine Learning

2.9.1 Application of articles whose keywords contain *machine learning*, *deep learning*, or *reinforcement learing*

- Autonomous Vehicles
  - Control of Autonomous Vehicles [5], [10], [21], [41], [44], [56], [68], [76], [80], [135], [165], [167], [172], [181], [189], [202],
  - Preferences for Autonomous Vehicles [193]
- Lagniappe
  - Anomalous Event Prediction [176],
  - Origin/Destination [96], [145],
  - Variable Speed Limits [172]
  - Dynamic Pricing [50], [59], [114]
  - Parking [100], [178], [195]
  - Traffic Signal Optimization [78], [86], [160], [163], [173], [182],
  - Perimeter Metering (?) [201]
  - Energy Consumption [120], [179]
  - Vehicle Idenfification [35], [83],

  - Trip Purpose [46]

- Traffic
  - Traffic Prediction [6], [11], [34], [37], [40], [42], [73], [84], [81], [82], [95], [130], [162], [170], [177], [196], [192],
  - Traffic Speed Prediction [106], [127], [157], [197]
  - Traffic in Extreme Weather [156]
  - Traffic Signals [57], [98], [198]
  - Dynamic Traffic Control [138]

- Individual Driver
  - Vehicle Behavior Modeling [29], [94], [102], [125], [180], [194]
  - Classifying Driving Styles [103]
  - Driver's Visual Environment [19], [85], [97]
  - Driver Behavior [103], [174],
  - Driver Distraction [19] This article is interesting, perhaps relevant to me, for correlating crashes with something else.

- Delivery
  - Delivery Times [63],
  - Vehicle Routing Problem [175], [191]
  - Fleet Management [152]
  - Transportation Systems [Survey article] [166]

- Public Transit
  - Taxis [27], [67], [70], [101], [121], [137], [149], [185],
  - Public Transit [31], [47], [90], [105], [147], [161], [158], [190]

- Pedestrians and Passengers
  - Pedestrians [18], [62]
  - Bicycles and Scooters [64], [93], [206],
  - Travel Demand Modeling [58], [72], [79], [89], [115]

- Planes, Trains, and Boats
  - Railway Maintenance [3]
  - Railway Traffic Control [51], [148]
  - Train Delays [87], [107]
  - Air Traffic Management [8], [2], [32], [38], [39], [43], [71], [111], [116], [119], [132], [153], [164], [204],
  - Ships [55], [92],

- Crashes
  - Inferring Pre-Crash Impact Data [28],

- Back End (No Application)
  - Generative Modeling [13], [49]
  - Preference Learning [203]
  - Extracting Economic Information (?) [159]
  - Graphs [129]
  - Discrete Choice Models [133]
  - Fairness in Artificial Intelligence [200]
  - Discrete Choice Modeling [171]

2.9.2 Articles whose abstracts refer to imbalanced data

Chen [29] talks about resampling using SMOTE and Tomek. Used LightGBM classifier.

Cai [20] used the deep convolutional generative adversarial network (DCGAN). Compared four models, logistic regression model, support vector machine, artificial neural network, and convolutional neural network.

Emarani Abou Elassad [45] works with several imbalanced methods. Use this paper as a model.

Yu [183] used focal loss for real-time crash prediction.

Shi [136] uses the Grey Wolf Optimizer and SMOTE to balance the data.

Khan [71] used SMOTE and “average balanced recall accuracies,”

Chen [30] uses bagging.

Anomalous events might also use imbalanced data.

2.9.3 Crashes

Twenty-one articles in TR_C have ’crash’, ’accident’, ’ambulance’, ’hospital’, ’fatal’, or ’injury’ in the keywords. Another forty have them in the abstract. I’m really only interested in ones that use real data, not simulation.

Kalatian [68] studies interactions between pedestrians and autonomous vehicles.

Cai and Abdel-Aty [20] do similar work to ours with machine learning.

Emarani Abou Elassad [45] was mentioned above as a model paper. Also applied to crashes.

Yu [183] mentioned above.

Chapter 3

Lit Review: Data Cleaning

3.1 Cleaning Techniques Used in Crash Analysis Studies

In “A deep learning based traffic crash severity prediction framework” by Rahim (LSU) [124], they just deleted any records with missing or inconsistent data. The *Titanic* Kaggle sites Dr. Jin showed me use several other methods for filling in incomplete data.

Rahim’s article took out 37% of the records for missing or inconsistent data, but only 21% of the fatal crashes; could that imbalance in the data cleaning skew the model prediction? It makes sense that police would be more meticulous in their record keeping for fatal crashes, but 21% and 37% are huge.

Chapter 4

Lit Review: Methods for Imbalanced Data

4.1 Algorithm Level Approaches

4.1.1 Some Papers

- Recognition-based: Learning from one class rather than discrimination-based, doing unsupervised learning on the minority class. [24]
- Fuzzy rule-based classification systems (what is this?) [22] [36] [99] [186] [188]
- In decision trees, using evolutionary/genetic methods instead of greedy search [22] [169]
- Clustering and Subspace Modeling [26]

4.1.2 Genetic Algorithms

In this short 2000 paper, Weiss [169] used a genetic algorithm to predict rare events. Borrowing from simulated annealing, they varied the relative importance of precision and recall at each step of the genetic algorithm.

4.1.3 Subspace Model

Chen 2011 [26]

This was fascinating and entirely different from anything I've seen.

1. Separate the training data Tr into negative (majority) and positive (minority) classes TrN and TrP.
2. Let K be the ratio of negative to positive samples, in my case about 100, so that if you divide the majority class TrN into K groups, each will have about the same number of samples as the minority class.
3. Use K-means clustering to separate the negative (majority) class TrN into K groups; each of the groups is a cluster of the negative (majority) class.
4. For each of the K groups TrN_i:
   - Combine the negative elements of the group with the entire positive (minority) class TrP to form a balanced subspace.
   - Train the model for the subspace
5. Recombine the K subspace models with a model trained on the entire data set to build an integrated model.

4.2 Metrics

4.2.1 The Problem: Imbalanced Data Set

In an unbalanced data set, the number of actual negatives ($N = TN + FP$) is much different from the number of actual positives ($P = FN + TP$). In our case, if our independent variable is fatal crashes, the negatives are 99.574714% of the data set, and the positives are just 0.425286%.

The standard metrics get thrown off by the imbalance. If we predict that every crash is nonfatal, we have accuracy of 99.57%, which sounds really impressive.

The recall (true positive rate) is not thrown off by an imbalanced data set, because it only works with TP and FN, the actual positives. Similarly for specificity (true negative rate).

The precision is thrown off by an imbalanced data set, because it works with both a subset of the actual positives (TP) and a subset of the actual negatives (FP).

4.2.2 Standard Metrics

		Prediction	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

		Prediction	
		N	P
Actual	N	TN	FP
	P	FN	TP

$$\begin{aligned}
\text{Accuracy} &= \frac{TN+TP}{TN+FP+FN+TP} \\
\text{Recall or TPR} &= \frac{TP}{TP+FN} \\
\text{Specificity, Selectivity, or TNR} &= \frac{TN}{TN+FP} \\
\text{Precision} &= \frac{TP}{TP+FP}
\end{aligned}$$

4.2.3 Balanced Precision and Balanced F1 in the Penalty Function

Most ML algorithms work using a *penalty function* that measures how bad the current solution is, then iteratively improving the solution in the direction that minimizes the penalty. We should be able to write a custom penalty function.

Update: How-to instructions for changing the metrics in `scikit-learn`. The example is how to use recall instead of accuracy.

`https://stackoverflow.com/questions/54267745`

Recall only deals with the minority class, so the balance of the data set doesn't matter. *Precision*, on the other hand, takes results from both classes, so we can balance it by scaling the count of False Positive results, giving a *Balanced Precision* metric. From Recall and Balanced Precision we can get a *Balanced f1* metric.

If our penalty function uses balanced precision and balanced f1, it may not matter that our data set is imbalanced, and we can use all of, and only, the original data to build our model.

4.2.4 Balanced Precision in the Literature

Balanced Accuracy frequently appears in the literature. I have not found *balanced precision* in the literature. Two possible reasons. Either nobody has thought of it, or they did, found it not useful, and abandoned the idea.

`imbalanced-learn` has more metrics than *scikit-learn*, but still no balanced precision.

`https://imbalanced-learn.org/dev/metrics.html`

4.2.5 Balanced Accuracy

There is a metric called *balanced accuracy.* You get it from the definition of *accuracy* by multiplying the actual negative elements (TN and FP) by the ratio of the positives to negatives,

$$\frac{P}{N} = \frac{FN + TP}{TN + FP}$$

so that the total number of actual negatives and total number of actual positives in the sample are equal.

[I suppose you could also get it by multiplying the actual positive elements (FN and TP) by the reciprocal.]

I got this derivation by intuiting about what I would want *balanced accuracy* to mean, and it matches the definition I found in Wikipedia.

`https://en.wikipedia.org/wiki/precision_and_recall#Imbalanced_data`

Wikipedia says [I'm sure I can find a more authoritative source.]

$$\text{Balanced Accuracy} = \frac{TPR + TNR}{2}$$

$$\begin{aligned}
\text{Recall or TPR} &= \frac{TP}{TP+FN} \\
\text{Specificity or TNR} &= \frac{TN}{TN+FP} \\
\text{Accuracy} &= \frac{TN+TP}{TN+FP+FN+TP} \\
\text{Balanced Accuracy} &= \frac{TN\cdot\frac{P}{N}+TP}{TN\cdot\frac{P}{N}+FP\cdot\frac{P}{N}+FN+TP} \\
&= \frac{TN\cdot P+TP\cdot N}{TN\cdot P+FP\cdot P+FN\cdot N+TP\cdot N} \\
&= \frac{TN\cdot P+TP\cdot N}{(TN+FP)\cdot P+(FN+TP)\cdot N} \\
&= \frac{TN(FN+TP)+TP(TN+FP)}{(TN+FP)(FN+TP)+(FN+TP)(TN+FP)} \\
&= \frac{TN(FN+TP)+TP(TN+FP)}{2(TN+FP)(FN+TP)} \\
&= \frac{TN(FN+TP)}{2(TN+FP)(FN+TP)}+\frac{TP(TN+FP)}{2(TN+FP)(FN+TP)} \\
&= \frac{TN}{2(TN+FP)}+\frac{TP}{2(FN+TP)} \\
&= \frac{TNR+TPR}{2}
\end{aligned}$$

4.2.6 Balanced Precision

I haven't found *balanced precision* in a brief Google search, although Google knows the kind of stuff I look up and sent me to articles on balanced accuracy. Finding it will take some work, because "balanced precision" has different meanings in other tech fields.

We can make balanced precision the same way we made balanced accuracy, by taking the actual negative results (TN and FP) and scaling them so that the total number of actual negatives equals the total number of actual positives, by multiplying by $\frac{P}{N}=\frac{FN+TP}{TN+FP}$.

Is this related to the G-mean? [No]

$$\text{G-mean} = \sqrt{\text{Precision} \times \text{Specificity}}$$

$$
\begin{aligned}
\text{Precision} &= \frac{TP}{TP + FP} \\
\text{Balanced Precision} &= \frac{TP}{TP + FP \cdot \frac{P}{N}} \\
&= \frac{TP \cdot N}{TP \cdot N + FP \cdot P} \\
&= \frac{TP(TN + FP)}{TP(TN + FP) + FP(FN + TP)} \\
&= \frac{TP(TN + FP)}{TP(TN + FP) + FP(FN + TP)} \\
&= \ldots
\end{aligned}
$$

Giving up here on finding some nice, concise connection between Balanced Precision and other metrics.

4.2.7 Balancing Two Metrics: F1 and Gmean

From Elassad 2020: [45]

> F1 score, is a highly informative measure as it considers both precision and recall measures, which makes it very suitable for imbalanced classification (Qian et al., 2014; Sun et al., 2018); it's deemed to be a special measure that conveys the balance between the precision and recall in order to find an effective and efficient trade-off. Another useful metric is G-mean, which is considered as a metric of stability between correct classification of positive class and negative class viewed independently. It is usually adopted in order to resist the imbalances in the dataset (Kubat et al., 1997).

F1 Metric

F1 is the harmonic mean of Precision and Recall.

$$
\text{F1} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}
$$

Gmean

Gmean is the geometric mean of Precision and Specificity (TNR).

$$\begin{aligned}\text{Precision} &= \frac{TP}{TP+FP}\\ \text{Specificity, Selectivity, or TNR} &= \frac{TN}{TN+FP}\\ \text{Gmean} &= \sqrt{\text{Precision} \times \text{Specificity}}\\ &= \sqrt{\frac{TP}{TP+FP} \times \frac{TN}{TN+FP}}\end{aligned}$$

4.3 Loss Functions

4.3.1 Binary Cross-Entropy Loss Function

Let's say we have an imbalanced data set with 100 negative samples for each positive sample.

For binary classification, the first three (class weights, weighted loss function, and naïve oversampling) are effectively the same in the training phase. The cross-entropy loss function,

$$loss = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$

for binary classification is

$$loss = \sum_{y_i=1} \log(p_i) + \sum_{y_i=0} \log(1 - p_i)$$

which is the sum of the logs of the errors in predictions for the negative class plus the sum of the logs of the errors in predictions for the positive class.

4.3.2 Class Weights and α-weighted Loss

If the classes are imbalanced, like there are 100 times as many samples with $y = 0$ as samples with $y = 1$, then the loss function is mostly summing how bad the predicting probability is for the majority class and largely ignoring the minority class. Both the class weights parameter and a weighted loss function fix this by multiplying one or the other by some compensating factor.

$$loss = 100 \times \sum_{y_i=1} \log(p_i) + \sum_{y_i=0} \log(1 - p_i)$$

This multiple gives the two classes equal weight in the loss.

In the α-weighted cross entropy,

$$loss = \sum_{i=1}^{n} \alpha y_i \log(p_i) + (1 - \alpha)(1 - y_i) \log(1 - p_i)$$

let $\alpha = \frac{100}{100+1}$ and you'll get the same thing, within a positive constant multiple.

$$loss = \sum_{i=1}^{n} \frac{100}{101} y_i \log(p_i) + \frac{1}{101}(1 - y_i)\log(1 - p_i)$$

$$loss = \frac{1}{101} \sum_{i=1}^{n} 100 y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

The only difference I can ascertain between the class weights parameter and a weighted loss function is that the class weights aren't used with the validation set.

4.3.3 Oversampling

4.3.4 Naïve Oversampling

Naïve oversampling would be to create 99 copies of each of the positive samples, so that the two sets are balanced. That would have exactly the same effect on the loss function, because there would now be 100 times as many samples with $y_i = 1$.

4.3.5 Class Weights v/s Naïve Oversampling: They're the Same

I had an insight on why these things are the same. Let's say you have an imbalanced data set, with 100 times as many negative samples as positive samples.

In Naïve Oversampling, you make 100 copies of each of the positive samples and run regular cross-entropy loss.

In weighted Class Entropy, you multiply the positive-class losses by 100.

$$loss = 100 \times \sum_{y_i=1} \log(p_i) + \sum_{y_i=0} \log(1 - p_i)$$

These two approaches different in execution but the same in result because, as I often remind my students, multiplying something by 100 is the same as adding it to itself 100 times.

4.3.6 Focal Loss

Introduced by Lin in 2017. [88]

Yu 2020 [184] adapts α-weighted cross entropy and focal loss to crash analysis.

In the focal loss function,

$$loss = \sum_{i=1}^{n} \alpha(1 - p_i)^{\gamma_1} y_i \log(p_i) + (1 - \alpha) p_i^{\gamma_2}(1 - y_i)\log(1 - p_i)$$
$$loss = \sum_{y_i=1} \alpha(1 - p_i)^{\gamma_1} \log(p_i) + \sum_{y_i=0} (1 - \alpha) p_i^{\gamma_2} \log(1 - p_i)$$

if $\gamma_1 = \gamma_2 = 0$, then it's the same as the α-weighted loss function.

In the original focal loss paper by Lin [88], γ_1 and γ_2 are the same.

For samples with $y_i = 1$, the minority class, here are values of $(1 - p_i)^{\gamma_1} \log(p_i)$ for different values of p_i and different values of γ_1. I got the range of values of $\gamma_1 \in \{0, 0.5, 1, 2, 5\}$ from Lin's 2018 paper that proposed focal loss.

		γ_1				
$(1 - p_i)^{\gamma_1} \log(p_i)$		0	0.5	1	2	5
	0.1	-3.32	-3.15	-2.99	-2.69	-1.96
	0.3	-1.74	-1.45	-1.22	-0.85	-0.29
p_i	0.5	-1	-0.71	-0.5	-0.25	-0.03
	0.7	-0.51	-0.28	-0.15	-0.05	0
	0.9	-0.15	-0.05	-0.02	0	0

If $\gamma_1 > 0$, then for samples in the positive class, the loss is negligible for good predictions (p_i close to 1), so it focuses the loss on poor predictions.

Yu applied focal loss in the crash literature.[184]

4.3.7 Optimizing F_β

Loss functions for gradient-based learning need to be differentiable (?), and the F_β score is not differentiable, so this 2021 article by Lee [77] proposes a differentiable surrogate loss function that optimizes the F_β score.

With imbalanced data, using a loss function that optimized F_β instead of accuracy would let you balance precision and recall, fixing one aspect of the imbalance problem.

$$F_\beta = \frac{(1+\beta^2) \cdot \text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}} = \frac{1}{\dfrac{\lambda_\beta}{\text{Recall}} + \dfrac{1-\lambda_\beta}{\text{Precision}}}, \qquad \lambda_\beta = \frac{\beta^2}{1+\beta^2}$$

The article takes a deep dive into loss functions. I should master it.

4.3.8 Tree-Based Methods

Pendault [118] has a 2000 article on insurance risk modeling that incorporates "a domain-specific optimization criterion... to identify suitable splits during tree building." It assigns different weights to *claim* and *nonclaim* records. Because that strategy helps but does not entirely solve the imbalanced data problem, they also have a split criterion that prevents splits of really small branches, "splinter groups," that are unlikely to contain any elements of the minority class because the minority class is so sparse.

4.3.9 α-weighted Binary Cross-Entropy Loss Function as Ethical Tradeoff

From `Brads_Report_10_25_21`

I made a [perhaps paper-worthy?] connection between the loss function I want and the α-weighted binary cross-entropy loss function, which is widely known and widely implemented, but, according to Yu's paper, not before used in crash-related modeling.

Matrix

	Do Not Send Ambulance $h_\theta(x_i) < 0.5$	Send Ambulance $h_\theta(x_i) > 0.5$
Do Not Need Ambulance $y_i = 0$	TN	FP
Need Ambulance $y_i = 1$	FN	TP

Switching between Binary and Continuous

In the binary cross-entropy loss function,

$$J = -\sum_{i=1}^{N} y_i \log(h_\theta(x_i)) + (1 - y_i) \log(1 - h_\theta(x_i))$$

the y_i are binary, $y_i \in \{0, 1\}$, but the model predictions, $h_\theta(x_i)$, are a probability, $h_\theta(x_i) \in (0, 1)$.

If we treat the model predictions as binary, replacing

$$\log(h_\theta(x_i)) \rightarrow \begin{cases} 0 & \text{if } h_\theta(x_i) <= 0.5 \\ 1 & \text{if } h_\theta(x_i) > 0.5 \end{cases}$$

and

$$\log(1 - h_\theta(x_i)) \rightarrow \begin{cases} 0 & \text{if } 1 - h_\theta(x_i) <= 0.5 \\ 1 & \text{if } 1 - h_\theta(x_i) > 0.5 \end{cases}$$

then

$$TP = \sum_{i=1}^{N} y_i \log(h_\theta(x_i))$$

$$TN = \sum_{i=1}^{N} (1 - y_i) \log(1 - h_\theta(x_i))$$

and the loss function becomes $J = -(TP + TN)$

Why do we use the continuous instead of the binary in the loss function? Because we want the predictions to be robust, so that when we use the model on unseen data, we can be more certain that it will correctly classify new instances. The binary, however, are much easier to explain to non-technical people, or even technical people in other fields.

Scenario

The medical ethicists and politicians decide on a number, p, such that we are willing to automatically dispatch p ambulances when they aren't needed in order to send one ambulance when it is needed. We want

$$\frac{\Delta FP}{\Delta TP} \leq p$$

Binary h_θ

Our loss function is

$$FP - p \cdot TP$$

Continuous h_θ

Use the α-weighted cross-entropy loss function, as in Yu's paper and widely available.

$$J = -\sum_{i=1}^{N} \alpha y_i \log(h_\theta(x_i)) + (1-\alpha)(1-y_i)\log(1-h_\theta(x_i)), \quad \alpha = \frac{p}{p+1}$$

Why are these equivalent?

Adding a constant to the loss function, or multiplying it by a positive constant, does not change its effect.

$FP - p \cdot TP$ is equivalent to $FP - p \cdot TP + (TN + FP)$, because $TN + FP$ is constant, so $FP - p \cdot TP$ is equivalent to $-(p \cdot TP + TP)$.

$$FP - p \cdot TP$$

$$-(p \cdot TP + TN)$$

Multiplying by $\frac{1}{p+1}$ gives an equivalent loss function, because $\frac{1}{p+1} > 0$.

$$-\frac{p \cdot TP + TN}{p+1}$$

$$-\left(\frac{p}{p+1}TP + \frac{1}{p+1}TN\right)$$

$$-\left(\frac{p}{p+1}TP + \left(1 - \frac{p}{p+1}\right)TN\right)$$

$$-(\alpha TP + (1-\alpha)TN)$$

The continuous version of TP is $\sum_{i=1}^{N} y_i \log(h_\theta(x_i))$

The continuous version of TN is $\sum_{i=1}^{N}(1 - y_i)\log(1 - h_\theta(x_i))$

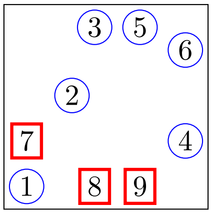
$$J = -\sum_{i=1}^{N} \alpha y_i \log(h_\theta(x_i)) + (1-\alpha)(1-y_i)\log(1-h_\theta(x_i)), \quad \alpha = \frac{p}{p+1}$$

Emphasis in Our Work

Yu et al introduced to the crash-analysis field the alpha-weighted cross-entropy loss function to deal with imbalanced data. We propose another application of the alpha-weighted loss, to encode and implement tradeoffs that come from our ethical/political values decided by community leaders.

4.4 Data Level Methods

Consider this two-dimensional training dataset, which we will use to illustrate data-level techniques for handling imbalanced datasets. In real problems, of course, the dataset could have a hundred dimensions and a million samples. The six blue circles represent samples (elements) of the majority negative class ("no ambulance"), and the three red squares represent the minority positive class ("ambulance").



Many algorithms, and variations thereon, have been proposed to balance the two classes before applying a machine learning algorithm to build a model to classify new samples as positive or negative. WARNING: Vast oversimplification ahead. Our goal here is to give the general idea of each method.

4.4.1 Imbalanced Cleaning: Tomek and Condensed Nearest Neighbor

Batista [9] uses two imbalanced cleaning method called *Tomek links* and *Condensed Nearest Neighbor*. If examples from the majority and minority class are close to each other, it deletes the majority samples. One could think of it as targeted undersampling of the majority set.

Imbalanced-Learn, an add-on to Scikit-Learn, has these algorithms read to use. Tomek and Wilson's papers introducing these algorithms are from the 1970's.

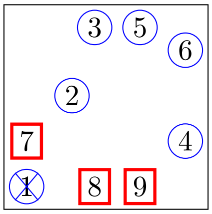
4.4.2 Tomek's Links

In 1976, Tomek proposed a method of undersampling that assumes that the majority and minority classes should (at least locally) be clustered. [150] If an sample A of the majority class and a sample B of the minority class are each other's nearest neighbors, then one of them is not clustered with its own class. Since we are trying to undersample the majority class, assume that the element of the majority class is noise (or an error, or just not useful), and delete it.

In the diagram below, samples #1 and #7 are Tomek links, because they are each other's nearest neighbors and of different classes. Samples #4 and #9 are not Tomek links, because while 9 is 4's nearest neighbor, 9's nearest neighbor is 8, not 4.

In the context of modeling crash severity from police reports, why would sample #1 not need an ambulance when its characteristics are so close to those of #7 and not near most of the other crashes without serious injury? The reason could be errors in the records, or luck/providence/fate. It could also be that the difference between property damage only and serious injury is influenced by thousands of variables we cannot measure or know, all of the physics of crash forces acting on the bones and structures of the human body. The best we can say is that the outcome in #1 cannot be predicted by the information that we have, so that sample will not help in constructing a model based on the available data; therefore, we can reasonably delete it from the training set.

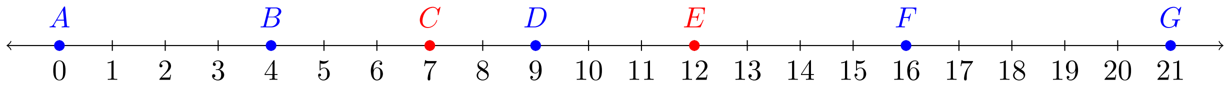
Tomek's Links can also be run iteratively. Sample #7 had #1 as its nearest neighbor, but once #1 is deleted, then #2 and #7 are each other's nearest neighbors of different classes, thus are Tomek links, and we can delete #2.



This method undersamples the majority-class samples, eliminating ones that are too close to minority-class samples, presuming them to be noise, and helping clarify clusters of minority samples.

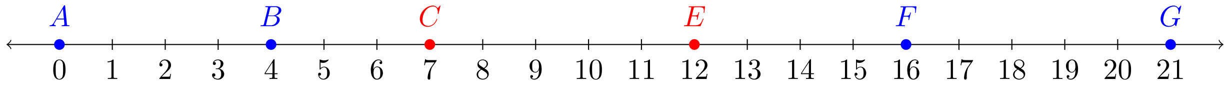
A pair of samples are a *Tomek link* if one is majority and one minority, and they are each other's nearest neighbors. To use Tomek's inks as an undersampling strategy for imbalanced data, delete the positive sample in each Tomek's link. Other cleaning strategies (for balanced sets) would eliminate both the positive and negative.

It is possible to iterate Tomek's several times. Here's an example of how it works in one round and in a second round. The blue samples are from the majority set and the red are from the minority. Assume that these seven points are a small part of a large dataset, but these are the only points in this region.

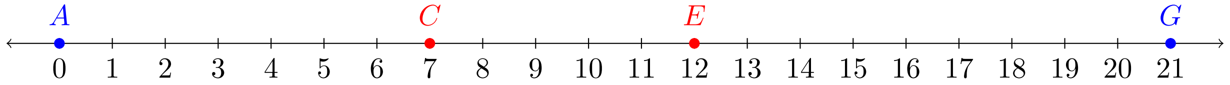


In the original dataset, C and D are each other's nearest neighbors, C from minority and D from majority, so they are a Tomek link. On the other hand, D is the nearest neighbor to E, but E is not D's nearest neighbor, so they are not a Tomek link.

Eliminate sample D.



Now the pairs (B, C) and (E, F) are Tomek links, so if we ran Tomek undersampling a second time, we would remove samples B and F.



Now C and E are each other's nearest neighbors and of the same (minority) class, so this part of the dataset would not change under another run of Tomek.

The idea of Tomek assumes that the minority samples should cluster, and any majority samples in or near those clusters must be noise, so we can eliminate them. We now have a clear cluster of two minority samples with no close majority samples.

I saw multiple runs of Tomek mentioned [somewhere] in my reading, so I tried it on the crash data, running it up to five times, and saw that it converged, with fewer positive samples eliminated in each round. I had conjectured that a negative sample in a Tomek link in a later round must have been a negative sample in a Tomek link in an earlier round, digging itself out of a field of positive-class dust, but I suspected that there might be (perhaps unusual) cases where one minority-class sample (C in the example above) created a Tomek link, and eliminating the majority-class sample in that link (D above) allowed a Tomek link for a different minority-class sample (E above). I then played with it until I found a counterexample to my conjecture, so the conjecture, that a minority-class sample in a Tomek link in a later round of Tomek undersampling must have been in a Tomek link in every previous round of the Tomek undersampling, is false.

If the conjecture had been true, then we could greatly speed up subsequent rounds of Tomek undersampling by only considering the minority samples in Tomek links in the previous round. That would not be thorough, but this approach would.

Algorithm for Repeated Application of Tomek's Links

For the first round of Tomek undersampling, one has to consider each element of the minority class. In the Tomek's links, call the minority-class elements $\{A_1, A_2, \ldots, A_{n1}\}$, and the majority-class elements $\{B_1, B_2, \ldots, B_{n1}\}$. Tomek undersampling for minority classes eliminates all of $\{B_1, B_2, \ldots, B_{n1}\}$.

In the second round of Tomek undersampling, we only need to consider as possible Tomek links the nearest neighbors of $\{A_1, A_2, \ldots, A_{n1}\}$ and any element of the minority class that had one of $\{B_1, B_2, \ldots, B_{n1}\}$ as its nearest neighbor.

In subsequent rounds, consider the minority-class samples from the Tomek's links from the previous round, and the elements of the minority class that had as their nearest neighbor an element of the majority class in the Tomek's links.

In theory there could be more Tomek's links in one round than in the previous round, but in practice they go to zero and the set converges to a set with no Tomek's links.

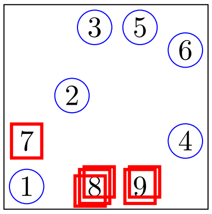
4.4.3 Cleaning Multiclass Data

Wei (2021) [168] uses something similar to Tomek's links for a multi-class problem with a majority class and multiple minority classes.

- Splits an imbalanced multi-class problem with $n + 1$ classes (n of them being minority) into n imbalanced binary problems for data cleaning.
- Uses cleaning undersampling (similar to Tomek's Links) to remove noisy spots in the data.

4.4.4 Random Oversampling

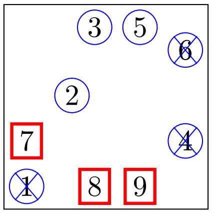
Random (naïve) oversampling creates duplicates of minority class samples until the sets are balanced. This method has a similar effect to using class weights, introduced below.



Naïve oversampling would be to create 99 copies of each of the positive samples, so that the two sets are balanced. That would have exactly the same effect on the loss function, because there would now be 100 times as many samples with $y_i = 1$.

4.4.5 Undersampling

Random undersampling balances the two classes by randomly deleting elements of the majority class until the two are balanced. The major drawback of this method is that you throw away information about the majority class. If the majority class is many more times the size of the minority, you lose almost all of the data.

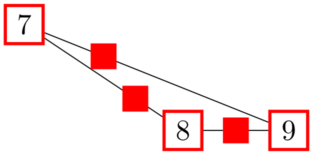


Undersampling would erase 99% of the negative samples so that the classes would be balanced. That seems like a bad idea, because you would lose a lot of information about the majority class.

4.4.6 SMOTE: Synthetic Minority Oversampling TEchnique

Synthetic Minority Oversampling Technique (SMOTE) [25] is one of the most popular oversampling methods for balancing a dataset with continuous numerical data. It creates new synthetic minority samples "between" original minority samples, not necessarily at the midpoint by choosing a number in $(0, 1)$, multiplying the difference (in each dimension) from point A to B by that constant, and adding it to A.

In the diagram, the solid red squares represent new synthetic samples between pairs of original minority-class samples. SMOTE does not consider the positions of the majority-class samples, only considering the difference in number of nodes to bring the two classes closer to parity.



One challenge with SMOTE is that it is only useful for datasets with continuous numerical data, and our data is almost all categorical. What is between "car" and "school bus," or between "parking lot" and "highway"? SMOTE has a variant, SMOTE-NC (Nominal and Continuous) that can handle datasets with some nominal (categorical) features, but most of the features need to be continuous; thus, we will not be able to use SMOTE or similar techniques for our work.

Especially if we're doing fatalities, we have a terribly imbalanced data set. Ideally we'd like to have an equal number of fatal and nonfatal crashes to plug into our ML algorithm, but we have about 0.47% fatal and 99.53% nonfatal.

One solution is to randomly choose 681 nonfatal crashes to compare with our 681 fatal crashes, but that leaves behind a LOT of information.

Many of the papers I've read use SMOTE, which balances the data set by creating synthetic elements for the minority set (fatal crashes). It picks an element of the minority set, a, and picks one of its nearest neighbors, b, and creates a new synthetic element c. For each data category, D_i, in which they differ, SMOTE chooses $D_i(c)$ to be between $D_i(a)$ and $D_i(b)$. It randomly chooses a random number $r \in [0, 1]$, and makes $D_i(c) = D_i(a) + r(D_i(a) - D_i(b))$.

I get how that works for continuous variables. I get that it would work if $D_i(a)$ and $D_i(b)$ weren't very different.

How would that work for boolean variables? SMOTE would choose nearest neighbors a and b that agree on most variables, but for values of i where $D_i(a) = 0$ and $D_i(b) = 1$, it would randomly choose $D_i(c) \in \{0, 1\}$. There is no *between* for boolean variables. It doesn't seem to me that it would work as well.

Original SMOTE only works with continuous variables. There is something called SMOTE-NC that handles continuous and categorical, but it has to have some continuous variables to work on.

> Unlike SMOTE, SMOTE-NC for dataset containing numerical and categorical features. However, it is not designed to work with only categorical features.

`https://imbalanced-learn.org/dev/references/generated/imblearn.over_sampling.SMOTENC.html`

Since we have ≈ 200 times as many nonfatal crashes as fatal crashes, to balance the data set with SMOTE, we would have to make two hundred synthetic elements for each fatal crash. It seems to me that we would be making a mess of our data set.

4.4.7 Flavors of SMOTE

SMOTE, or Synthetic Minority Oversampling TEchnique, [23] creates extra samples of the minority class, but rather than making exact copies, it finds two similar samples and creates more samples "between" them, with feature values between the values of the two samples. SMOTE only works for continuous features, not for categorical features. Almost all of my features are categorical.

I got this list of flavors of SMOTE from a 2021 review by Mahmudah. [99] I've investigated some of them and given some flesh to some parts of this skeleton.

- SMOTE: Synthetic Minority Oversampling TEchnique [23]
  Uses k-nearest neighbors to find two close positive (minority) samples, and creates a synthetic sample between them. Works on continuous data, not on categorical or binary data.
- ADASYN: ADAptive SYNthetic sampling approach for imbalanced learning. [99]
  Creates synthetic samples based on the level of difficulty in learning the samples of the minority class. A positive samples is "difficult" if it has more negative samples as its nearest neighbors. The more difficult a sample is, the more synthetic copies of that sample ADASYN creates.
- Borderline SMOTE [99]
  Generates synthetic positive samples along the border between the positive and negative classes. Brad's Question: This assumes you know where the border is. I suppose you could do it iteratively.
- Safe-level SMOTE [99]
  When SMOTE finds the nearest positive-class neighbors of a positive sample, it ignores the negative (majority-class) neighbors. [I think this is what it means:] Creating synthetic positive-class samples in a neighborhood with lots of negative samples just makes more of a mess, so this is not considered a "safe" place to make synthetic samples. Safe-level SMOTE creates synthetic positive samples only in majority-positive neighborhoods.
- Relocating-safe-level SMOTE (RSLS) [99]
  Avoids creating synthetic positive samples near negative samples.
- Density-based SMOTE (DBSMOTE) [99]
  Integration of DBSCAN and SMOTE. DBSCAN, Density-Based Spatial Clustering of Application with Noise, discovers clusters with an arbitrary shape (?) DGSMOTE creates synthetic samples at the pseudo-centroids of the clusters of positive samples.
- Adaptive Neighbor SMOTE (ANS) [99]

Focuses not on -where- to generate synthetic samples, but on -how many- samples to generate in a particular neighborhood.

- D2GAN

  This 2020 article by Zhai [188] builds on the Dual Discriminator Generative Adversarial Nets (D2GAN) paper from 2017 by Nguyen [109]. They want to do better oversampling, comparing D2GAN with SMOTE. I don't understand what this is, but they say SMOTE has three drawbacks:

  1. Ignores the probability distribution of minority class samples.
  2. Synthetic examples lack diversity.
  3. Interating SMOTE many times will give synthetic samples with significant overlap.

  This 2022 article by Zhai [187] slightly modifies Zhai's claims against SMOTE.

  1. Does not extend the training field of positive samples.
  2. Synthetic examples lack diversity.
  3. Does not accurately approximate the probability distribution of minority class samples.

  The authors propose two new methods of diversity oversampling by generative models, one based on "extreme machine learning autoencoder," and the other based on generative adversarial networks (GAN).

4.4.8 Oversampling Image Data

Extracting knowledge from a database of tabular numerical or categorical data is difficult, but a database of images is a challenge of a different magnitude. An imbalanced labeled image dataset for crash prediction modeling might be a thousand images taken ten seconds before a crash and a million images taken ten seconds before ... nothing happened. Deep neural networks (DNN) and (deep) convolutional neural networks (DCNN and CNN) are common methods for image data. [54] introduced Generative Adversarial Networks, which can be used to generate synthetic samples to balance the dataset. Given the power of the tools for image recognition, many researchers make non-image data look (to the computer) like images to take advantage of the tools.

4.4.9 Train/Test Split

The application in Sharififar's 2019 article [134] is digital mapping of farmland, categorizing areas by soil type. Some soil types are rare but significant. This is the first article I've seen that, at the beginning, says that making sure each minority class appears in appropriate distribution in the validation and test sets is an important challenge. They explicitly say that they split 30% for the validation set by taking 30% of each class.

4.4.10 Feature Selection

This 2012 article by Tan [146] introduces a feature selection model specifically for imbalanced data sets. I haven't dug in yet.

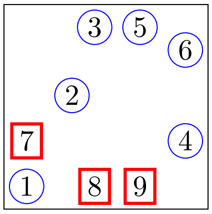
4.5 Bagging and Boosting

Bagging

"Bagging" is short for Bootstrapped Aggregating, a variation on random undersampling. [15] In general, bagging takes many random subsets (with replacement) of the samples, run the classifier on each subset, then aggregate the results. In imbalanced data applications, each subset of the samples is all of the n minority samples and n randomly chosen majority samples.

Balanced Random Forest [need citation] is a form of bagging.

In our example, bagging would make a subset of the data with the three minority-class samples (#7, 8, and 9), and three randomly chosen from the majority-class samples, run the classifier; repeat some number of times. Use an ensemble classifier to merge the results.



Lack (2021) used bagging in predicting crashes for trucks and finding ways to improve truck safety. [74]

Shi (2021) developed a hierarchical over-sampling bagging method based on Grey Wolf Optimizer (GWO) algorithm and Synthetic Minority Over-sampling Technique (SMOTE) to study lane changing for autonomous vehicles. The data was severely imbalanced because lane changing is rare compared with lane keeping. [136]

Chen (2022) used bagging for ride-hailing demand prediction. [30]

Boosting

Boosting is an iterative method that runs the classifier multiple times. At the end of each iteration, it determines which samples would be misclassified under the current model. In the next iteration, the classifier gives higher weight to the misclassified samples, improving the model on marginal cases. While boosting is not just for imbalanced data, the challenge in imbalanced data is that the minority class samples get misclassified, so boosting would help. A popular implementation is AdaBoost, introduced by [48].

Haule (2021) used boosting in studying the effects of ramp metering on traffic safety. [60]

- Boosting and Bagging [9] [22] [36] [99] [134]

4.6 Lit Review: Medium.com *Towards Data Science* Articles

These aren't exactly peer reviewed, but they're current.

Soleymani (4/1/22) says that class weights are more effective than SMOTE, and gives an example of why SMOTE doesn't do what you think it should. [140]

Raj (9/5/19) is a brief article that introduces what an imbalanced data set is, and resampling, including naïve oversampling, undersampling, and SMOTE. [126]

Soni (10/9/20) introduces Balanced Random Forest, with code, in addition to undersampling and oversampling. Balanced Random Sampling is, I think, a form of bagging. You take a bootstrap sample of the minority class and the same number of elements from the majority class, and run random forest; then aggregate the results. [141]

Brownlee isn't in TDS, but gives an easy introduction to ROC curves. [17] Also gives good references in [16].

Stewart also mentions Tomek Links. [143]

Bordia reviews variants of SMOTE, including SMOTE_NC, which works with datasets with some (but not all) categorical data and some continuous data. NC is for Nominal and Continuous. [12]

Boyle recommends Random Forests for imbalanced data. [14]

Keras can do random forest classifiers, although you may need to make it yourself. `https://keras.io/examples/structured_data/deep_neural_decision_forests/`

How to do an ROC curve and find AUC for Keras and sklearn: `https://medium.com/hackernoon/simple-guide-on-how-to-generate-roc-plot-for-keras-classifier-2ecc6c73115a`

Badr includes bagging. [7]

Rocca gives many different ideas. Read this one carefully. [128]

Lador gives good examples of when different metrics are useful. [75]

Jaitley also recommends Random Forest, Gradient Boosting, and AdaBoost. [65]

Ahamed had entirely different recommendations, Ensemble Cross-Validation (CV), Class Weights, and Over-Predicting the class of the minority class, *i.e.* setting a lower probability threshold for the minority class. [1]

Chapter 5

Lit Review: Datasets

5.1 Crash Datasets

- SHRP2, Strategic Highway Research Program 2, Naturalistic Driving Study
  Federal Department of Transportation
  Most cited dataset.
- Second Highway Research Program (Data Set)
  I have an account.
- Virginia 100-car Database
- Next Generation Simulation, NGSIM Trajectory Data https://iswitrs.chp.ca.gov/Reports/jsp/index.jsp
- NASS-CDS: National Automotive Sampling System – Crashworthiness Data System
- Canada's National Collision Database
- Michigan Safety Pilot
- Roadway Information Database (RID)
- Shanghai Naturalistic Driving Study
- California Statewide Integrated Traffic Records System (SWITRS)
  Apparently anyone can get an account?
  `https://iswitrs.chp.ca.gov/Reports/jsp/index.jsp`
- Highway Safety Information System
  Not updated since 2018?
  `http://www.hsisinfo.org`

5.1.1 Jargon to Understand

From `24_May_2021_Report`:

- Naturalistic Driving Data - Data collected from sensors installed in the driver's own car, trying to get as close as possible to the driver's "natural" behavior.
- Heterogeneity. I understand vaguely what "data heterogeneity" means, but I'm going to watch for the term to see how it's used in the context of these papers.

5.1.2 IRB, SHRP Database

Eleven of the papers in *Accident Analysis and Prevention* used the Strategic Highway Research Program 2 (SHRP2) Naturalistic Driving Study (NDS), which put sensors in 3400 cars and recorded five million trips, including crashes. To get "Qualified Researcher Status" with "full access to data that has been made available through the SHRP 2 NDS Data Access Website," I had to submit a certificate of training on research with human subjects. I did the training through the UL Institutional Review Board (IRB). I now have access.

5.1.3 NGSIM Database

Three papers use the Next Generation Simulation dataset from the US Dept of Transportation, and it's available for download with no restrictions.

5.2 Datasets with Imbalanced Data

5.2.1 Datasets, Annotated

5.2.2 Articles using These Datasets

Zheng 2021 [199]

- Oversampling, undersampling, and hybrid methods use random sampling ratios. [What? How? I thought the user set the sampling ratios.]
- This paper proposes three algorithms to automatically set the sampling ratios using genetic algorithms.
- Used fourteen datasets, some of which may be useful benchmark datasets.

Wang 2021 [155]

- Uses seven benchmark imbalanced datasets from the UCI machine learning repository
- Implicit regularization for dynamic ensemble selection of classifiers.

5.2.3 Database Repositories

UCI Machine Learning Repository

`https://archive.ics.uci.edu/ml/about.html`

Chapter 6

Lit Review: Seminal and Interesting Papers

6.1 Seminal Papers

- Lin [88] introduced Focal Loss in 2017. The 2017 versions of this article are only available through Inter Library Loan, because the UL Library apparently doesn't subscribe to IEEE, and the version I found was from 2020.

6.2 Review Papers

6.2.1 Chawla

Chawla [24] gives an overview of the state of the field in 2004.

- Data Methods
  - Random Oversampling with Replacement
  - Random Oversampling
  - Directed Oversampling
    No new examples are created, but the choice of which ones to replace is informed rather than random.
  - Directed Undersampling
  - Oversampling with informed generation of new samples
  - Combinations of the above
- Algorithmic Methods
  - Adjusting class costs
  - Adjusting the probabilistic estimate at the tree leaf (for tree methods)
  - Recognition-based methods (learning from one class) rather than discrimination-based.
- Issues at 2000 Conference
  -
- Issues at 2003 ICML Conference
  - Probabilistic estimates
  - Pruning
  - Threshold adjusting
  - Cost-matrix adjusting.
- Interesting Topics at 2003 ICML Conference
  - Selective sampling based on query learning (Abe)
- Overlapping Problems
  - Class Imbalance
  - Small Disjunct Problem (?)
  - Rare Cases

  - Data Duplication
  - Overlapping Classes

By 2003, the field started to mature.

6.2.2 Chabbouh 2019

This article [22] has a nice table classifying existing work in imbalanced classification; however, I think much of the information was old in 2019, particularly C4.5, an early decision tree base classifier that may not be used much anymore.

6.2.3 Mahmudah 2021

This article [99] is really a review of current methods. They have some datasets, most public benchmark sets, and throw every combination of tools at them. The "methods" section is really an overview of current methods.

Has a section on techniques for feature extraction (feature engineering?) by dimensionality reduction, not particularly related to imbalanced data.

6.3 Examples of Good Writing, Models to Follow

- Elassad 2020 [45] is a good model.
- Paez 2021 [113] is not ML, but a solid paper. The conclusion suggests looking into imbalanced learning.
- Soleimani (LSU) 2019 [139] gives a thorough analysis.

6.3.1 Elassad 2020

Good model to follow.

- In the title and first sentence of the abstract talks about an application, Collision Avoidance Systems, that the paper does not work with directly, which is like what I'm doing with mobile phones.
- Has several glaring mistakes, like crash avoidance systems on the vehicle having access to data from loop detectors, which are embedded under the road.
- Projects into the future, assuming that vehicles will detect the physiological state of the driver. I do this when I assume that police departments will have access to up-to-date and well-calibrated maps, to personal data from phone companies, and to be able to corollate several pieces of data (from multiple phones) in real time.
- Critique: Doesn't define terms well. What is an "ensemble fusion framework"? How are "ensemble" and "fusion" different? In layman's language, they sound the same. Uses "fusion" to mean both classifier ensembles and data fusion.

- Good overview at the end of the Introduction.
- The ML guts of this paper are trying different combinations of classifiers for an ensemble method. The guts of my paper will be different combinations of imbalanced data techniques.
- Only uses two imbalanced data techniques: Class weights and SMOTE.
- Algorithms
- Table of features
- Six points for future research

Chapter 7

CRSS Dataset

7.1 CRSS Overview

The Crash Report Sampling System (CRSS) [110] is from the National Highway Transportation Safety Board (NHTSB), part of the US Department of Transportation (DOT). Available data is from 2016-2020. In 2016, CRSS replaced the National Automotive Sampling System General Estimates System (NASS GES), which goes back to the 1970's.

> The CRSS obtains its data from a nationally representative probability sample selected from the more than six million police-reported crashes that occur annually. To be eligible for the CRSS sample, a crash report must be completed by the police; it must involve at least one motor vehicle traveling on a trafficway; and the crash must result in property damage, injury, or death.
>
> These crash reports are chosen from 60 selected sites across the United States that reflect the geography, population, miles driven, and crashes in the United States. CRSS data collectors review crash reports from hundreds of law enforcement agencies within the sites, systematically sampling tens of thousands of crash reports each year. The collectors obtain copies of the selected crash reports and send them to a central location for coding. No other data is collected beyond that in the selected crash reports.
>
> Trained personnel interpret and code data directly from the crash reports into an electronic data file. Approximately 120 data elements are coded into a common format. After coding, quality checks are performed on the data to ensure validity and consistency. When these are completed, CRSS data files and coding documentation become publicly available. [108]

The data comes with a helpful user's manual [108] and a guide to their imputation of missing values that includes a history going back to the 1980's. [61]

7.2 CRSS Data Files

Each year's CRSS dataset comes in twenty-some .csv files, but most are derivatives of the main three, ACCIDENT, VEHICLE, and PERSON, and from henceforth I will only mention these three.

The term "accident" has fallen out of favor, because it implies that the crash was not intentional, by commission or omission, so the practitioners in the field prefer "crash." CRSS and the journal *Accident Analysis and Prevention* may keep "accident" for historical consistency. I will tend to use "crash," except when referring to the ACCIDENT data file.

Each accident in ACCIDENT has a case number, CASENUM, and has at least one corresponding vehicle in VEHICLE. One can merge the two sets on the case number. Each accident has at least one vehicle, and each vehicle belongs to an accident.

Each sample in VEHICLE has a vehicle number, VEH_NO, numbered from 1 in each accident. In PERSON, each sample has the case number of the accident. If the person was in a vehicle, then the sample has the vehicle number. If the person was not in a vehicle, for instance a pedestrian, then the vehicle number is 0. Not all vehicles have a person, and not all persons have a vehicle, so merging the two datasets requires handling values that are properly blank.

For our work, we dropped all crashes with pedestrians, because the deceleration profile of a crash between a vehicle and a pedestrian, on the phones of the pedestrian or an occupant of the vehicle, is different from the deceleration profile of hitting another vehicle or a tree. The deceleration profile would be so similar to hard braking that we doubt the phone would send an alert.

7.3 CRSS Features (178 Features)

7.3.1 Imputed Features to Use ($10 \times 2 = 20$ Features)

Notes

- The RELJCT1 field did not have missing values imputed in 2019, so those 54,409 cells are blank in RELJCT1_IM. To reconstruct it, I used the RELJCT1_IM values from 2016, 2017, 2018, and 2020 with the RELJCT1 values from 2019, and used IVEware to impute the missing values. [122] Not perfect, but better than we had. See the CRSS Imputation report for details. [61]
- Why did CRSS impute missing values for DAY_WEEK when there weren't any missing values? For historical consistency and backwards compatibility going back to 1988. [61] In a crash report, some data may be missing because of human error, confusion, rush, illegibility, . . . , but the date is one of the first things on the report and is more reliable.

Original Feature	Imputed Feature	Meaning	Number of Categories	Num. of Missing Values	Values Signifying "Unknown"	Num. of Unknown Values
AGE	AGE_IM	Age	118	0	[998,999]	41087
BODY_TYP	BDYTYP_IM	Vehicle Body Type Code	73	0	[98, 99, 49, 79]	18211
DAY_WEEK	WKDY_IM	Day of Week	7	0	[9]	0
HOUR	HOUR_IM	Hour	25	0	[99]	1127
LGT_COND	LGTCON_IM	Light Condition	9	0	[8,9]	2309
MOD_YEAR	MDLYR_IM	Model Year	83	0	[9998, 9999]	18524
RELJCT1	RELJCT1_IM	Relation to Junction-Within Interchange Area	4	54409	[8,9]	65920
RELJCT2	RELJCT2_IM	Relation to Junction-Specific Location	15	0	[98,99]	19721
SEX	SEX_IM	Sex	4	0	[8,9]	26143

WEATHER	WEATHR_IM	Weather	13	0	[98,99]	13284

7.3.2 Features to Use with No Missing or Unknown Values (13 Features)

Feature Name	Meaning	Number of Categories
MODEL	Vehicle Model Code	140
MONTH		12
PEDS	Number of persons not in motor vehicles	10
PER_TYP	Person type	13
PERMVIT	Number of Persons in Motor Vehicles in Transport	26
PERNOTMVIT	Number of Persons Not in Motor Vehicles in Transport	10
PVH_INVL	Number of Parked/Working Vehicles in the Crash	11
REGION		4
SCH_BUS		2
URBANICITY		2
VE_FORMS	Number of Motor Vehicles in Transport	13
VE_TOTAL	Number of vehicles in crash	13
WRK_ZONE	Work Zone	5

7.3.3 Features to Use with Unknown Values to Impute (12 Features)

Feature Name	Meaning	Number of Categories	Num. of Missing Values	Values Signifying “Unknown”	Num. of Unknown Values
HOSPITAL	How taken to hospital	9	0	[8,9]	13522
INT_HWY	Interstate Highway	3	0	[9]	25
MAKE	Vehicle Manufacturer Code	70	0	[99]	12901
MOD_YEAR	Model Year	83	0	[9998, 9999]	18524
REL_ROAD	Relation to Trafficway	13	0	[98,99]	190
TYP_INT	Type of Intersection	11	0	[98,99]	26650
VALIGN	Roadway Alignment	7	0	[8, 9]	31554
VNUM_LAN	Total Lanes in Roadway	10	0	[8, 9]	127387
VPROFILE	Roadway Grade	9	0	[8, 9]	62776
VSPD_LIM	Speed Limit	20	0	[98, 99]	62649
VTRAFCON	Traffic Control Device	19	0	[97, 99]	30151
VTRAFWAY	Trafficway Description	9	0	[8, 9]	83513

7.3.4 CRSS Internal Features for Merging the ACCIDENT, VEHICLE, and PERSON Data Files (3 Features)

Note that if the person was not in a vehicle, PER_NO = 0.

Feature Name	Meaning	Number of Unique Values
CASENUM	CRSS Case Number	259077
VEH_NO	Index of Vehicle in Crash	15
PER_NO	Index of Person in Vehicle	75

7.3.5 CRSS Imputed Features to Not Use, Except as a Control for our Imputation Method ($17 \times 2 = 34$ Features)

These features are unknowable without investigation on the scene, thus not relevant to our study of features that are either given by an automated report or can be inferred from one.

These fields, however, like the CRSS-imputed fields above, may be useful as a control for our imputation method, which is an approximation of the method used by the CRSS authors. We can impute the unknown values in the original feature and compare our imputations those from CRSS.

Original Feature	Imputed Feature	Meaning	Number of Categories	Num. of Missing Values	Values Signifying “Unknown”	Num. of Unknown Values
ALCOHOL	ALCHL_IM	Alcohol Involved in Crash	4	0	[9]	59889

DRINKING	PERALCH_IM	Person Drinking	4	0	[8,9]	232366
EJECTION	EJECT_IM	Ejection	7	0	[9]	2137
HARM_EV	EVENT1_IM	First Harmful Event	56	0	[98,99]	166
HIT_RUN	HITRUN_IM	Hit and Run	3	94718	[9]	30
IMPACT1	IMPACT1_IM	Area of Impact – Initial Contact Point	26	0	[98, 99]	11061
INJ_SEV	INJSEV_IM	Injury Severity	8	0	[9]	21595
M_HARM	VEVENT_IM	Most Harmful Event	56	0	[98, 99]	189
MAN_COLL	MANCOL_IM	Manner of Collision of the First Harmful Event	11	0	[98,99]	1012
MAX_SEV	MAXSEV_IM	Maximum Severity in Crash	9	0	[9]	4480
MAX_VSEV	MXVSEV_IM	Maximum Injury Severity in Vehicle	9	0	[9]	18600
MINUTE	MINUTE_IM		61	0	[99]	1127
NUM_INJ	NO_INJ_IM	Number Injured in Crash	20	0	[99]	4480
NUM_INJV	NUMINJ_IM	Number Injured in Vehicle	17	0	[99]	18600
P_CRASH1	PCRASH1_IM	Pre-Event Movement (Prior to Recognition of Critical Event)	20	0	[99]	8340
SEAT_POS	SEAT_IM	Seating Position	30	0	[98,99]	7981
VEH_ALCH	V_ALCH_IM	Driver Drinking in Vehicle	4	0	[9]	84494

7.3.6 Other Features to Not Use (94 Features)

We excluded these features for one or more of these reasons.

- Not knowable without on-scene investigation, so cannot even be guessed well from a cell phone notification.
- Useless information, like registration number or license number.
- Data in that feature is not available for all five years.

7.4 CRSS Binning

Model building is more efficient and effective if the number of categories in each feature is reasonably small, with “reasonably” being fuzzy, but ten is a good target. If some of the categories are essentially the same, it is better to bin (merge) them together, especially if some of the categories are very small.

In the Crash Report Sampling System (CRSS) data set, all of the features we plan to use are categorical, and most have a small number of categories. The features for Age, Vehicle make, Vehicle model, Model year, and Vehicle body type each have more than fifty categories. Some of them (age, model year) are ordered, and the rest are not. To identify "similar" categories, I looked at how each category correlated with the target variable, being taken to a hospital.

First I binned the HOSPITAL feature into a binary feature. A few steps later I will get to imputing unknown values, but at this stage I binned the "Not Reported" and "Reported as Unknown" in the vastly majority category, "Not Transported."

To Do: Put titles on tables.

To Do: Standardize the horizontal and vertical spacing of tables.

Original Code	Bin	Number of Samples	Meaning
0	0	522,801	Not Transported
1	1	2,549	EMS Air
2	1	605	Law Enforcement
3	1	30,368	EMS Unknown Mode
4	1	8,926	Transported Unknown Source
5	1	61,162	EMS Ground
6	1	4,341	Other
8	0	12,447	Not Reported
9	0	1,075	Reported as Unknown
All	0	536,323	83.24%
	1	107,951	16.76%

Then for each value in AGE, I found the percentage of samples with that value and the percentage of samples of that value who were transported to a hospital. A part of the results is in the table below. Note the big shifts between ages 14, 15, 16, and 17, perhaps suggesting that new drivers are prone to more fender benders but not serious crashes, and that we should make $[15, 16]$ its own bin.

	Value	Percent of Samples with this Value	Percent with this Value Hospitalized
AGE_IM	10	0.53	16.58
AGE_IM	11	0.51	15.51
AGE_IM	12	0.52	16.54

AGE_IM	13	0.54	16.72
AGE_IM	14	0.63	17.56
AGE_IM	15	0.88	15.21
AGE_IM	16	1.65	13.46
AGE_IM	17	2.18	14.25
AGE_IM	18	2.65	14.41
AGE_IM	19	2.67	15.47
AGE_IM	20	2.58	14.91

A similar shift in the hospitalization rate occurs in the early 50's, so we made another split between 51 and 52.

	Value	Percent of Samples with this Value	Percent with this Value Hospitalized
AGE_IM	44	1.39	15.23
AGE_IM	45	1.36	16.08
AGE_IM	46	1.33	15.69
AGE_IM	47	1.39	15.26
AGE_IM	48	1.34	16.46
AGE_IM	49	1.32	16.59
AGE_IM	50	1.29	16.64
AGE_IM	51	1.39	16.24
AGE_IM	52	1.36	16.57
AGE_IM	53	1.24	17.05
AGE_IM	54	1.31	17.74
AGE_IM	55	1.28	17.66
AGE_IM	56	1.23	17.21

Using the correlation to hospitalization rates, we set the bins for AGE at $[0, 14]$, $[15, 16]$, $[17, 51]$, and $[52, \infty)$.

The same technique was especially useful with BODY_TYP. The table below shows the major body types ($\geq 0.40\%$ of samples). The horizontal lines show where we divided the CRSS categories into bins.

CRSS Value	Description	Percent of Samples with this Value	Percent with this Value Hospitalized

BODY_TYP	80	Two Wheel Motorcycle (excluding motor scooters)	2.11	61.48
BODY_TYP	2	2-door sedan,hardtop,coupe	2.92	15.66
BODY_TYP	3	3-door/2-door hatchback	0.73	15.46
BODY_TYP	1	Convertible(excludes sun-roof,t-bar)	0.62	15.29
BODY_TYP	30	Compact Pickup (Only used in 2016)	0.40	15.26
BODY_TYP	4	4-door sedan, hardtop	33.87	15.24
BODY_TYP	19	Utility Vehicle, Unknown body type	0.99	14.57
BODY_TYP	5	5-door/4-door hatchback	2.38	14.00
BODY_TYP	49	Unknown light vehicle type (automobile,utility vehicle, van, or light truck)	2.08	13.95
BODY_TYP	8	Sedan/Hardtop, number of doors unknown	0.71	13.93
BODY_TYP	6	Station Wagon (excluding van and truck based)	4.87	13.42
BODY_TYP	14	Compact Utility (Utility Vehicle Categories ”Small” and ”Midsize”)	14.46	13.26
BODY_TYP	9	Other or Unknown automobile type	2.92	12.79
BODY_TYP	20	Minivan (Chrysler Town and Country, Caravan, Grand Caravan, Voyager, Voyager, Honda-Odyssey, ...)	3.90	12.48
BODY_TYP	34	Light Pickup	8.66	11.30
BODY_TYP	31	Standard Pickup (Only used in 2016)	1.57	10.88
BODY_TYP	15	Large utility (ANSI D16.1 Utility Vehicle Categories and ”Full Size” and ”Large”)	4.83	10.76
BODY_TYP	39	Unknown (pickup style) light conventional truck type	0.50	9.95
BODY_TYP	21	Large Van-Includes van-based buses (B150-B350, Sportsman, Royal Maxiwagon, Ram, Tradesman,...)	1.10	8.74
BODY_TYP	61	Single-unit straight truck or Cab-Chassis (GVWR range 10,001 to 19,500 lbs.)	0.62	5.26
BODY_TYP	67	Medium/heavy Pickup (GVWR greater than 10,000 lbs.)	0.42	4.55
BODY_TYP	63	Single-unit straight truck or Cab-Chassis (GVWR greater than 26,000 lbs.)	0.46	4.25
BODY_TYP	66	Truck-tractor (Cab only, or with any number of trailing unit; any weight)	1.62	3.50

Based on the hospitalization rates, we binned the seventy-three CRSS categories into five.

New Bin	CRSS Value
Motorcycle	[86,87,82,89,81,83,80,84,88,85,90,95,11,97,96,58,45,12,32,91]
Car	[10,2,3,1,59,30,4]
SUV	[19,42,5,49,8,16,6,14,52,9, 20]
Light Truck	[22,40,34,31,15,29,92,39,55,93,17,21,50,48,28,7,65]
Heavy Truck	[51,61,67,63,62,66,79,78,64,72,98,60,99,71,73,94,41,13]

Most of the groupings we could have done by name, putting the cars together, the SUV's together... But some names do not fit their hospitalization profile, like "Compact Pickup,"

> Compact Pickup (S-10, LUV, Ram 50, Rampage, Courier, Ranger, S-5, Pup, Mazda Pickup, Mitsubishi Truck, Datsun/Nissan Pickup, Arrow Pickup, Scamp, Toyota Pickup, VW Pickup, D50, Colt P/U, T-10, S-15, T-15, Ram 100, Dakota, Sonoma)

which has the hospitalization profile of a small car, not a light truck.

We binned other features similarly.

7.5 CRSS Imputing Unknown Values

The Crash Report Sampling System (CRSS) is the latest iteration of National Highway Transportation Safety Board (NHTSB) datasets, and for historical consistency and backwards compatibility, the CRSS authors only imputed unknown values in some features. The CRSS authors wrote a very useful historical and practical report on their imputation methods. [61]

7.5.1 "Missing" v/s "Unknown"

I will distinguish here between "missing" and "unknown" data. In each year's CRSS spreadsheets, no cells are blank, but some (less important) features appear in one year and not another, so when I merge them I get an entire year of blank cells. I will refer to those as "missing."

When I merge the Vehicle and Person parts of each year's data, the vehicle data will be blank or `nan` for some samples because the person was not in a vehicle (pedestrian, bicyclist, motorist parked on the side of the road and standing outside the vehicle...) Those samples I have dropped for reasons described earlier.

7.5.2 Unknown within Bin v/s Unknown Unknown

Almost all of the features are categorical, and most of them include at least one category signifying that the value is unknown. Sometimes they are partially unknown but contain enough information for our purposes, like in the HOSPITAL feature, in the table below. Category 6 is "Other," which is undefined. To see what kinds of severity it covers, we look at the crosstabs with injury severity

(INJ_SEV) (next table below). Category 6 looks similar to category 5, “EMS Ground” in that all of the people had some injury, and most of the injuries were minor. (See the INJ_SEV / HOSPITAL Crosstabs Normalized by Row table below). Thus, we interpret “Other” as “Transported to hospital by another means.” We actually don’t care how the person was transported to the hospital, just whether the person went, so we will bin 1, 2, 3, 4, 5, and 6 together.

Categories 8 and 9 of HOSPITAL are unknown unknown. One method of handling an unknown category is to bin it in the largest bin, but we used a more subtle method, using IVEware [122].

HOSPITAL

Category			Percentage
	Meaning	Count	of Samples
0	Not Transported	522801	81.15
1	EMS Air	2549	0.40
2	Law Enforcement	605	0.09
3	EMS Unknown Mode	30368	4.71
4	Transported Unknown Source	8926	1.39
5	EMS Ground	61162	9.49
6	Other	4341	0.67
8	Not Reported	12447	1.93
9	Reported as Unknown	1075	0.17

INJ_SEV: Injury Severity

0	No Apparent Injury
1	Possible Injury
2	Suspected Minor Injury
3	Suspected Serious Injury
4	Fatal Injury
5	Injured, Severity Unknown
6	Died Prior to Crash
9	Unknown/Not Reported

INJ_SEV / HOSPITAL Crosstabs

INJ_SEV	0	1	2	3	4	5	6	9
HOSPITAL								
0	429574	52271	19522	826	2956	454	9	17189
1	0	159	295	1854	222	16	0	3
2	0	266	235	84	7	3	0	10
3	0	9630	10601	8801	665	618	3	50
4	0	3293	2686	2476	226	215	1	29
5	0	22550	19983	16642	1315	470	6	196
6	0	2214	1546	419	91	31	0	40
8	0	5106	2308	1421	95	72	0	3445
9	0	272	123	33	10	4	0	633

INJ_SEV / HOSPITAL Crosstabs Normalized by Row (%)

INJ_SEV	0	1	2	3	4	5	6	9
HOSPITAL								
0	82.17	10.00	3.73	0.16	0.57	0.09	0.00	3.29
1	0.00	6.24	11.57	72.73	8.71	0.63	0.00	0.12
2	0.00	43.97	38.84	13.88	1.16	0.50	0.00	1.65
3	0.00	31.71	34.91	28.98	2.19	2.04	0.01	0.16
4	0.00	36.89	30.09	27.74	2.53	2.41	0.01	0.32
5	0.00	36.87	32.67	27.21	2.15	0.77	0.01	0.32
6	0.00	51.00	35.61	9.65	2.10	0.71	0.00	0.92
8	0.00	41.02	18.54	11.42	0.76	0.58	0.00	27.68
9	0.00	25.30	11.44	3.07	0.93	0.37	0.00	58.88

7.5.3 IVEware

The CRSS Features (178 Features) section above listed twenty features (that we want to use) whose unknown values had been imputed by the CRSS authors and another twelve features, like HOSPITAL, whose unknown values had not been imputed. The CRSS Imputation report describes the reasons why some features were imputed and other not, mainly for historical consistency going back to 1988. [61] As best we could, we replicated their methods for the twelve features with unknown values.

1. Impute unknown values in ACCIDENT dataset
2. Merge VEHICLE into ACCIDENT
3. Impute unknown values in VEHICLE
4. Merge in PERSON
5. Impute missing values in PERSON

The CRSS authors used Imputation and Variance Estimation Software (IVEware) to implement

Sequential Regression Multivariate Imputation (SMRI) for their first round of imputing unknown values. [122] [123] They wrote a very useful report on their methods, with an historical overview and the hyperparameters they used when running IVEware. [61] The authors followed up the SMRI with manual updates based on domain knowledge, but only of twenty-eight samples. We will not be able to replicate that part of their method.

7.5.4 IVEware Testing

When the CRSS authors imputed unknown values for a feature, they left the unimputed feature in the dataset. To test how close our imputation results were to theirs, we ran our imputation algorithm on some of those unimputed features and compared. Since imputation involves randomness, we also compared two of our imputation runs to see whether the difference between our results and those of CRSS were largely due to expected random variability and not significantly due to a difference in methods.

The tables below compare imputation results for the Lighting Conditions feature, LGT_COND, just looking at the 2,309 samples with unknown values 8 or 9.

The tables below are for Lighting Conditions. The LGT_COND feature is the original data with 2,309 unknown values. The LGTCOND_IM is the imputed feature in the CRSS data set, and the LGT_COND_IVE is one run of our imputation. The LGT_COND_IVE_2 is our imputation with the same hyperparameters but a different random seed. The crosstabs below only show the 2,309 samples with unknown values, comparing the values to which the imputations assigned them.

LGT_COND (Light Conditions)

Value	Meaning	Count
1	Daylight	177,013
2	Dark - Not Lighted	26,403
3	Dark - Lighted	41,508
4	Dawn	4,063
5	Dusk	6,016
6	Dark - Unknown Lighting	1,697
7	Other	68
8	Not Reported	1,690
9	Reported as Unknown	619

To Do: Include totals for rows and columns

Comparing CRSS's Imputation with Ours

LGT_COND_IVE	1	2	3	4	5	6	7
LGTCON_IM							
1	946	57	52	42	22	8	1
2	67	381	132	37	25	21	1
3	55	114	161	7	6	9	0
4	25	15	6	11	0	1	0
5	21	35	17	0	6	3	0
6	5	9	5	0	2	3	0
7	0	0	1	0	0	0	0

Comparing Two IVEware Runs with Different Random Seeds

LGT_COND_IVE_2	1	2	3	4	5	6	7
LGT_COND_IVE							
1	936	80	48	27	21	7	0
2	72	364	108	24	21	21	1
3	60	108	187	2	11	6	0
4	40	25	6	25	0	1	0
5	20	25	12	1	0	3	0
6	9	26	6	2	1	1	0
7	1	1	0	0	0	0	0

To Do : Find a metric for comparing the variability.

The two crosstabs tables indicate that, while my recreation of the imputation method used by the CRSS authors does not give the same results, it gives similar results, and the differences are largely consistent with the differences between two runs with different random seeds. Our imputation is not the same, but may be as similar as possible, given that we are working with unknowns and randomness.

7.6 CRSS Binning and Imputing: Order of Operations

The HOSPITAL feature has nine values, one representing that we know the person was not transported to a hospital, six representing that we know the person was transported, and two representing that we don't know.

We will want to impute those unknowns and bin the feature into a binary with 0 for Not Transported and 1 for Transported.

HOSPITAL Feature

Original Value	Bin	Number of Samples	Meaning
0	0	522,801	Not Transported
1	1	2,549	EMS Air
2	1	605	Law Enforcement
3	1	30,368	EMS Unknown Mode
4	1	8,926	Transported Unknown Source
5	1	61,162	EMS Ground
6	1	4,341	Other
8		12,447	Not Reported
9		1,075	Reported as Unknown

Before we do the binning and imputing, we need to decide which to do first. Do we impute unknown values 8 and 9 into {0, 1, 2, 3, 4, 5, 6}, then condense {1,2,3,4,5,6} into {1}, or do we condense {1,2,3,4,5,6} into {1}, then impute 8 and 9 into {0,1}? Would they give the same results? Would we be able to determine how much of the difference is due to randomness in the imputation algorithm?

For some features we had to do the binning first, because the feature had too many categories for IVEware to handle. The feature MAKE has 70 values, and MOD_YEAR has 83. Some experimentation and conversations with IVEware staff showed that fewer that forty categories was possible. The CRSS Imputation report describes how the CRSS authors did that with AGE, putting it into bins by decade, imputing missing values, then putting back the known values. [61]

I asked Dr. Raghavan which should come first, binning or imputing, and he told me that the answer depends on the data, and I should experiment. He sent me a paper on the topic by one of this students. I plan to experiment and decide which to use. My gut hypothesis is that binning first would be better, but we will not see a definitive difference in the test results.

To Do: Investigate Order of Operations for binning and imputation.

7.7 CRSS Feature Engineering (Rough)

Binning a feature is a form of feature engineering, but here I mean merging elements from two or more features into a new feature.

One new feature we have created is a binary feature Rush Hour, combining Day of Week (WKDY_IM) and Hour (HOUR_IM), using the percent of people hospitalized in crashes at a particular hour on a weekday to draw the lines.

We also looked at crossing AGE_IM with SEX_IM to bin ages separately by gender, because the correlation to hospitalization is more complicated than either feature separately.

7.8 CRSS in the Literature

- Torpuz and Delen (2021) [151] does a thorough description of imputing missing data in CRSS. Does not mention IVEware. Also deals with imbalanced data well. Need to spend time with this article.
- Cox and Cicchino (2021) [33] says CRSS "can be weighted to produce annual national estimates." Also, "Police-reported crash sampling methods changed when NHTSA converted from NASS GES to CRSS, which may have affected the comparability of the 2017 data on all crash involvements with earlier years."

  In this study, "Imputed data were utilized when available to account for missing data."
- Amini, Bagheri, and Delen (2022) [4] gives a thorough description of CRSS. They took out CRSS-imputed variables. Also removed post-accident information, as it was not relevant. They imputed missing continuous variables, but don't say how. They left missing categorical variables as "Unknown" and "Missing" categories.

  > Employing descriptive analytics, we distinguished and removed variables with a large percentage of missing values (more than 70%), as well as the identification, irrelevant, repetitive, and CRSS-imputed variables. We also removed the variables with post-accident information, such as whether the vehicle was towed afterward or the number of injured people. Using such variables contradicts the basic assumption of time order in causal relations, where a cause should precede its effect. Furthermore, we handled other missing values by considering them separate categories for nominal variables and imputing numeric ones.

- Spicer et al (2021) [142] used CRSS but did not mention missing or imputed data.
- Villavicencio, Svancara, Kelly-Baker, and Tefft (2022) [154] says that "CRSS is a representative sample of all police-reported crashes in the United States," which is not true. They used FARS and CRSS as their primary data sources, but did not mention imputed or missing data.
- Mueller and Cicchino (2022) [104] says that "The CRSS data set handles missing data for some variables by statistically imputing values, which were used when available."
- Kaplan et al [69] uses the phrase, "restricted access database." I should use that for the Louisiana crash database.
- Gong et al [53] just dropped samples with missing values.
- As far back as 2002, NHTSA was working on multiple imputation methods for its related database, FARS. [144]

Chapter 8

Louisiana Dataset

8.1 Overview

The Louisiana dataset, a census of crash reports, has restricted access, and I only have a portion of it. I cannot give readers access to the data to check or build on my work; thus, I am focusing on the CRSS data but using the Louisiana data for another viewpoint and to get experience with a different kind of data.

I have the data 2014-2018. Its organization is similar to that of CRSS, with Crash, Vehicle, and Occupant datasets to be merged. The Crash data set for those five years has 828,248 samples.

I have not worked with the Louisiana database since March, when I changed my focus to the CRSS database. When I go back to the Louisiana database, my approach will be different. Before, I had only used the Crash data set, not the Vehicle and Person data sets; I will use all three, as I did with the CRSS. When I started with the Louisiana, I just looked at fatalities, but now I will look at hospitalization.

8.2 Properties

8.2.1 Boolean Nature of our Data

Most of our data is boolean. Was alcohol involved? Did the car leave its lane? Was there a pedestrian? We have categorical variables, like type of vehicle which we represent as dummy (boolean) variables. We have some categories we could represent as numbers (like day of the week), and we could impose an order, (Monday comes before Tuesday), but the order isn't relevant in predicting injuries or fatalities, (Neither increases or decreases as the days "progress."), so we should represent them as categories, in dummy variables.

8.2.2 Top Twenty Features that Correlate with Fatality

Last column is the *balanced f1* score.

DR_COND_CD2	I	DRUG USE - IMPAIRED	0.33
SEC_CONTRIB_FAC_CD	L	CONDITION OF PEDESTRIAN	0.32
PRI_CONTRIB_FAC_CD	L	CONDITION OF PEDESTRIAN	0.25
PRI_CONTRIB_FAC_CD	M	PEDESTRIAN ACTIONS	0.20
VEH_TYPE_CD1	G	OFF-ROAD VEHICLE	0.18
M_HARM_EV_CD1	B	FIRE/EXPLOSION	0.17
DR_COND_CD2	F	APPARENTLY ASLEEP/BLACKOUT	0.17
CRASH_TYPE	C	[Unknown]	0.17
SEC_CONTRIB_FAC_CD	M	PEDESTRIAN ACTIONS	0.16
M_HARM_EV_CD1	O	PEDESTRIAN	0.15
VEH_COND_CD	E	ALL LIGHTS OUT	0.15
F_HARM_EV_CD1	O	PEDESTRIAN	0.15
M_HARM_EV_CD1	F	FELL/JUMPED FROM MOTOR VEHICLE	0.15
F_HARM_EV_CD1	F	FELL/JUMPED FROM MOTOR VEHICLE	0.14
PEDESTRIAN			0.13
VEH_TYPE_CD1	E	MOTORCYCLE	0.13
DR_COND_CD2	G	DRINKING ALCOHOL - IMPAIRED	0.13
CRASH_TYPE	A	[Unknown]	0.13
MOVEMENT_REASON_2	G	VEHICLE OUT OF CONTROL, PASSING	0.12

8.3 Thoughts on our Data Set: Trees

I suspect that a decision tree is the only realistic way to make a predict model for any aspect of crash data. If a pedestrian is involved, or it's a rural area, or alcohol is involved, the dynamics of the problem change. That there could be some linear (or nonlinear) function of all of the variables to fatality or injury is not reasonable to hope. If we think of it not as one big problem but as lots of little problems, like "What factors predict a fatality/injury in a crash involving a pedestrian in a rural area at night?" and, "What factors predict a fatality/injury in a crash where alcohol is involved at rush hour in an urban area?", we'll have much more likelihood of success.

8.4 Times

From the `Brads_Report_11_01_21`

8.4.1 New Features

Interesting features I didn't have before:

- `AMBULANCE` $\in \{0, 1\}$
- `CRASH_TIME`
- `TIME_POLICE_NOTE`

- `TIME_POLICE_ARR`
- `TIME_AMB_CALLED`
- `TIME_AMB_ARR`

In the 828,248 records, 167,662 (20.2%) have `AMBULANCE`==1.

8.4.2 Misspellings

In the 'CITY' feature in the data, the name of the city of Shreveport is spelled nineteen different ways. It's not a problem, though, because it's spelled correctly about 47,000 times and incorrectly only 35 times.

8.4.3 Dirty Data

In many of the records, one of the times could be 0, which could indicate midnight, but more likely indicates missing data. Lots of the records mix up AM and PM. Some of them have the police or ambulance called before the crash time. In some of them, the ambulance isn't called until more than half an hour after the crash time, which could be real, but more likely a data entry error. Adding to the messiness is that some of the crashes roll over midnight.

There may be ways to fix some of those records, but for now I'll thrown them out. I threw out 47,640 records (28%), leaving 120,002 records.

8.4.4 Strange Data

The `CRASH_TIME` feature is in the format "1/1/01 HH:MM:SS," but the second are either "00" or "39." I don't know why. I'm going to ask Malek whether it appears that way in the original Access file.

8.4.5 Ambulance Call within/after 5 min after Crash

- In 64% of the cases, the ambulance was called within 5 min of the crash.
- In 15% of the cases, the ambulance was called more than 5 min after the crash and after the police arrived. Those 18,037 are the interesting cases.

8.4.6 Hospitalized

Of the 120,022 clean records where an ambulance was called, 43,902 (37%) had no one hospitalized, so while the ambulance crew may have applied minor first aid, it wasn't an emergency.

Chapter 9

Methods and Experimental Results To Date (Louisiana Dataset)

9.1 scikit-learn

I ran just about every scikit-learn classifier, with results in my `12_July_2021_Report`.

Most Keras examples I see use tools from scikit-learn as well.

There's an add-on to scikit-learn called imbalanced-learn which has SMOTE, Tomek, and other tools.

9.2 Focal Loss and Tomek

Working with our crash database, with the cleaning and organizing in which I had it in February 2022, I tried different values for γ_1 and γ_2 with and without Tomek Links cleaning.

Tomek Links is a method for cleaning a noisy dataset for binary classification. A *Tomek Link* is a pair of samples, one from the positive and one from the negative class, that are each others' closest neighbors. The idea is that one of them is noise, or that having these two interferes in making a good classification, that you want the classes to cluster. In a balanced dataset you eliminate both of them from the training set. In an imbalanced dataset, you eliminate the element from the majority class.

From my weekly report 2/21/22:

- Unfortunately, p means two different things below.
  - p_i is the probability returned by the model that each sample belongs to the positive set.
  - p is a hyperparameter, ideally the proportion of the negative to positives samples, to use $\alpha = p/(p+1)$ in the Focal Loss function, to create the class weights that have the same effect as random oversampling. In our dataset, $p = 88.8$. (No, I'm not kidding.)
- All runs without Tomek used the same training and test sets
- All runs with Tomek used the same training and test sets
- The two test/train splits used the same random sampling seed, so they should be the same sets.

$$\begin{aligned}\text{Focal Loss} &= \sum_{i=1}^{n} \alpha(1-p_i)^{\gamma_1} y_i \log(p_i) + (1-\alpha)p_i^{\gamma_2}(1-y_i)\log(1-p_i)\\ &= \sum_{y_i=1} \alpha(1-p_i)^{\gamma_1}\log(p_i) + \sum_{y_i=0}(1-\alpha)p_i^{\gamma_2}\log(1-p_i)\end{aligned}$$

9.2.1 Different Values of p with $\gamma_1 = 0$, $\gamma_2 = 0$

Tomek?	p	γ_1	γ_2	TN/FN	FP/TP	Comments
No	1	0	0	573308	0	
				6466	0	
No	20	0	0	562182	11126	
				5850	616	
No	88.8	0	0	428929	144379	This is the natural p
				3105	3361	for our dataset.
No	100	0	0	411813	161495	
				2737	3729	
No	200	0	0	287151	286157	
				1464	5002	

9.2.2 Fixed $p = 88.8$, Different values of γ_1 and γ_2

Tomek?	p	γ_1	γ_2	TN/FN	FP/TP	Comments
No	88.8	0.0	0.0	428929	144379	This is the natural p
				3105	3361	for our dataset.
No	88.8	0.5	0.5	399870	173438	
				2685	3781	
No	88.8	1.0	1.0	420343	152965	
				3092	3374	
No	88.8	2.0	2.0	433805	139503	
				3213	3253	
No	88.8	5.0	5.0	445445	127863	
				3519	2947	
No	88.8	0.0	2.0	337148	236160	
				2092	4374	
No	88.8	0.5	0	460148	113160	
				3520	2946	
No	88.8	1.0	0.0	391820	181488	
				2596	3870	
No	88.8	2.0	0.0	527871	45437	
				4877	1589	

9.2.3 Tomek

Tomek took out 760 negative samples, bringing p down to 88.66.

Tomek?	p	γ_1	γ_2	TN/FN	FP/TP	Comments
No	88.8	0.0	0.0	428929	144379	
				3105	3361	
Yes	88.8	0.0	0.0	387313	185995	FN goes up 29%
				2504	3962	TP goes up 18%
Yes	88.66	0.0	0.0	387313	185995	
				2504	3962	

9.2.4 Discussion

- The different values of p, γ_1, and γ_2, and Tomek, give us different tradeoffs between false positives and false negatives, but no combination gives us fewer of both.
- It would be challenging to argue that one set of hyperparameters is "better" than another.
- I suspect that there just isn't enough of a pattern in this crash data to give us much confidence.
- I need to also work on other datasets that either might give clearer results, or will show me that all results are this fuzzy and I need to learn how to deal with it.

9.3 Feature Engineering

9.3.1 Time of Day

Time of day is a continuous variable, but the correlation between time of day and [anything] is nonlinear. We could do some kind of data transformation, perhaps taking the ratio of the number of accidents to the typical traffic density at that time of day, but the typical car trip at 3 am on a Wednesday may be different in character than a car trip at 7 am on a Saturday, even if the traffic volumes are similar. Perhaps we should have boolean variables:

- Morning rush hour
- Mid-day
- Afternoon rush hour
- Evening
- Late night

and another variable, `Weekend`.

9.3.2 Number of Fatalities/Injuries

The number of fatalities or injuries is a function of how many people were in each vehicle, which (a) we don't know and (b) probably isn't correlated to any other data we have. Fatality and injury should be boolean variables, that there was a fatality or there was an injury, rather than a count of the number of fatalities or injuries.

9.3.3 Day of Week

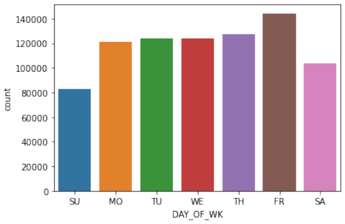


Figure 9.1: Number of Crashes, by Day of Week

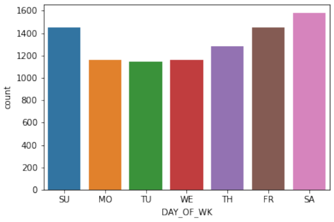


Figure 9.2: Number of Severe Injury Crashes, by Day of Week

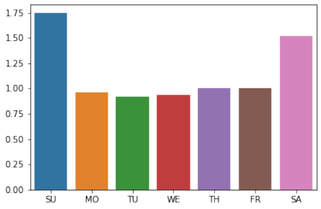


Figure 9.3: Percentage of Crashes with Severe Injury, by Day of Week

My understanding is that, for feature engineering, we don't care that there are more crashes on Friday than other weekdays, since the proportion of crashes that require an ambulance are the same. Saturday and Sunday, though, are different.

I made a feature, `Weekday_SA_SU`:

0	MO, TU, WE, TR, FR
1	SA
2	SU

9.3.4 Time of Day

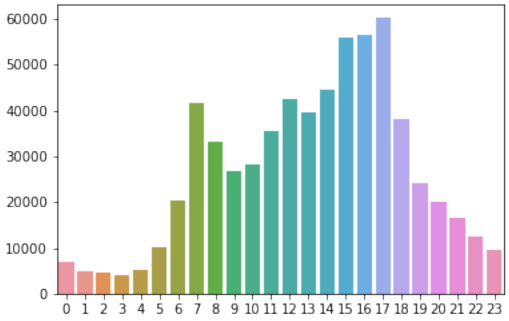


Figure 9.4: Number of Weekday Crashes, by Hour

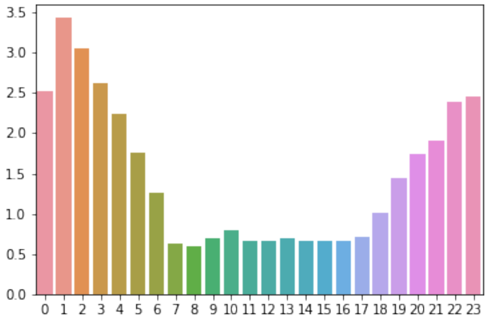


Figure 9.5: Percentage of Weekday Crashes with Severe Injury, by Hour

I note with interest that, at 7am, the number of crashes spikes, but the percentage of severe injury crashes does not change significantly. I created a `Rush_Hour` feature, but I don't know if it will be of any use.

The spike of percentage of crashes at 1am is just noise, because of the small number of crashes at that time.

The types of roads on which crashes occurs varies widely by time of day. I don't know what to do with that.

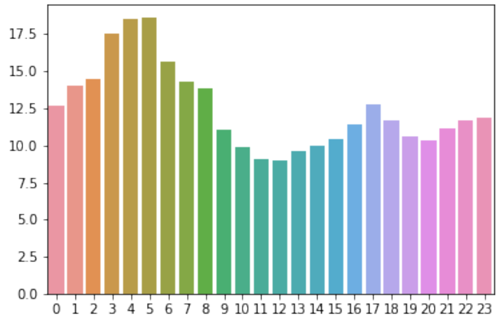


Figure 9.6: Percentage of Crashes on Interstates, by Hour

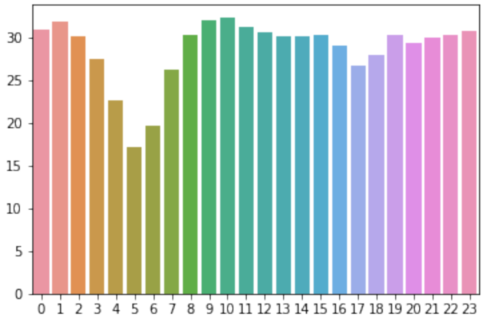


Figure 9.7: Percentage of Crashes on City Streets, by Hour

I made a feature, `Time_of_Day`, grouping together times with similar percentages of crashes having severe injuries:

0	Midnight - 3:00 am
1	3:00 am - 5:00 am
2	5:00 am - 7:00 am
3	7:00 am - 5:00 pm
4	5:00 pm - Midnight

9.3.5 Location

Location seems like it would be very important. One proxy we have is the parish and the road name. I've made a new feature, concatenating the parish and the road name. Each unique value in that feature will become a category, yielding a new feature in the dummy (one-hot encoding) dataframe that we will use for training.

There are 6.150 unique values, but most of them have few records. How many records do you need to make a useful correlation, and how many categories will overload the training?

Having a minimum of 1000 records per category gives me 142 categories plus 492,367 records in “Other”; a minimum of 100 records gives me 1103 categories plus 221,644 records in “Other.” A minimum of 10 records gives me 2,534 categories and 171,802 in “Other.” Note that 161,454 are in “Other” because of missing data.

9.3.6 Parish/Road Names

- We have 161,454 records with ”0” for the `PRI_ROAD_NAME`. There's nothing we can do to recover those.
- We have 26,289 different values for `PRI_ROAD_NAME`.
- We have even more if we combine those with the, sometimes multiple, `PRI_ROAD_TYPE`, like St, Ave, and Blvd.
- In a few instances, roads with the same `PRI_ROAD_NAME` and different `PRI_ROAD_TYPE` are different roads, but usually within the same parish they're the same. A noteable exception is North St and North Blvd in Baton Rouge.
- For long roads, like interstates and some state highways, crash outcomes may differ based on which section of road you're on.
- To Do:
  - Combine `PARISH_CD` and `PRI_ROAD_NAME` into a new feature, `PARISH_CD_and_PRI_ROAD_NAME`.
  - Ignore the `PRI_ROAD_TYPE`
  - Keep the instances of `PARISH_CD_and_PRI_ROAD_NAME` that have more than 1000 crashes.
  - Change all of the others to ”Other”.
- Results:
  - This leaves us with 142 different names with 335,880 crashes, plus “Other” with 492,367 crashes.

- `PRI_ROAD_NAME` = ”AIRLINE” appears (with at least 1000 crashes) in 11 parishes, with 51,399 crashes.

Chapter 10

Research Plan

10.1 Progress To Date

- Reviewed literature (ongoing process)
- Developed problem
- Chose datasets (Crash Report Sampling System (CRSS) [110] and Louisiana)
- Learned how to build custom loss functions in Keras. (It's not really an option in scikit-learn.)
- Understood a wide variety of methods for handling imbalanced data. Many of them are available in Keras, and some I had to implement myself.
- Learned to use Imputation and Variance Estimation Software (IVEware) [122] and used it to impute unknown values in CRSS

10.2 Goals

10.2.1 CRSS Data Set

Crash Report Sampling System (CRSS) [110]

- Answer question about whether binning or imputing should come first.
- Some of the binning is not consistent; make a clear rationale for binning and apply it.
- Finish preparing the data.
- Apply imbalanced data techniques, testing individually and in combination.
- Analyze results.

10.2.2 Paper for *Transportation Research Part C: Emerging Technologies*

- Reread papers from this journal that are models of good writing.
- Read and reread the submission policies.
- Revise.
- Submit.
- Get feedback.
- Respond to feedback.

10.2.3 Louisiana Data Set

- Select features to match/complement what I did with CRSS data.
- Clean.
- Discretize data. This will be different from what I did with CRSS, because some of the data is continuous.
- Impute missing values.
- Apply imbalanced data in a way that matches/complements what I did with the CRSS data.
- Analyze results.

10.2.4 Write Dissertation

- Review the literature again
- Write
- Revise
- Repeat

10.3 Timeline

October 2022	Answer question for CRSS about order of operations of binning and imputing unknown values Finish preparing CRSS data
November 2022	Test imbalanced data techniques (and combinations thereof) on CRSS data
December 2022	Analyze results
January 2023	Submit paper to *Transportation Research Part C: Emerging Technologies*
February 2023	Clean Louisiana database Respond to reviews from TR_C
March 2023	Wrestle with the data: Figure out how to use Louisiana and CRSS data together
April 2023	Test imbalanced data techniques (and combinations thereof) on the Louisiana data
May 2023	Read, Write, and Revise
June 2023	Read, Write, and Revise from Spain and France
July 2023	Read, Write, and Revise from Turkey
August 2023	Read, Write, and Revise
September 2023	Read, Write, and Revise
October 2023	Submit Dissertation
December 2023	Dissertation Defense
15 December 2023	Graduation

Bibliography

[1] Sabber Ahamed. *Three Important Techniques to Improve Learning Model Performance with Imbalanced Datasets*. 2018. URL: https://towardsdatascience.com/working-with-highly-imbalanced-datasets-in-machine-learning-projects-c70c5f2a7b16 (visited on 07/01/2022).

[2] Berat Mert Albaba, Negin Musavi, and Yildiray Yildiz. “A 3D game theoretical framework for the evaluation of unmanned aircraft systems airspace integration concepts”. In: *Transportation Research Part C: Emerging Technologies* 133 (2021), p. 103417. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103417. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004113.

[3] Zaharah Allah Bukhsh et al. “Predictive maintenance using tree-based classification techniques: A case of railway switches”. In: *Transportation Research Part C: Emerging Technologies* 101 (2019), pp. 35–54. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18309057.

[4] Mostafa Amini, Ali Bagheri, and Dursun Delen. “Discovering injury severity risk factors in automobile crashes: A hybrid explainable AI framework for decision support”. In: *Reliability Engineering & System Safety* 226 (2022), p. 108720. ISSN: 0951-8320. DOI: https://doi.org/10.1016/j.ress.2022.108720. URL: https://www.sciencedirect.com/science/article/pii/S0951832022003441.

[5] Yunlong An et al. “Space-time routing in dedicated automated vehicle zones”. In: *Transportation Research Part C: Emerging Technologies* 120 (2020), p. 102777. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102777. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306872.

[6] Danya Bachir et al. “Inferring dynamic origin-destination flows by transport mode using mobile phone data”. In: *Transportation Research Part C: Emerging Technologies* 101 (2019), pp. 254–275. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.02.013. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18310519.

[7] Will Badr. *Having an Imbalanced Dataset? Here Is How You Can Fix It.* 2019. URL: https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb (visited on 07/01/2022).

[8] Jie Bao, Zhao Yang, and Weili Zeng. "Graph to sequence learning with attention mechanism for network-wide multi-step-ahead flight delay prediction". In: *Transportation Research Part C: Emerging Technologies* 130 (2021), p. 103323. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103323. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003296.

[9] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data". In: *SIGKDD Explor. Newsl.* 6.1 (June 2004), pp. 20–29. ISSN: 1931-0145. DOI: 10.1145/1007730.1007735. URL: https://doi.org/10.1145/1007730.1007735.

[10] Rolando Bautista-Montesano et al. "Autonomous navigation at unsignalized intersections: A coupled reinforcement learning and model predictive control approach". In: *Transportation Research Part C: Emerging Technologies* 139 (2022), p. 103662. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103662. URL: https://www.sciencedirect.com/science/article/pii/S0968090X2200105X.

[11] Toon Bogaerts et al. "A graph CNN-LSTM neural network for short and long-term traffic forecasting based on trajectory data". In: *Transportation Research Part C: Emerging Technologies* 112 (2020), pp. 62–77. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.01.010. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19309349.

[12] Ansh Bordia. *Handling Imbalanced Data by Oversampling with SMOTE and its Variants.* 2022. URL: https://medium.com/analytics-vidhya/handling-imbalanced-data-by-oversampling-with-smote-and-its-variants-23a4bf188eaf (visited on 07/01/2022).

[13] Stanislav S. Borysov, Jeppe Rich, and Francisco C. Pereira. "How to generate micro-agents? A deep generative modeling approach to population synthesis". In: *Transportation Research Part C: Emerging Technologies* 106 (2019), pp. 73–97. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.07.006. URL: https://www.sciencedirect.com/science/article/pii/S0968090X1831180X.

[14] Tara Boyle. *Dealing with Imbalanced Data.* 2019. URL: https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18 (visited on 07/01/2022).

[15] Leo Breiman. In: *Machine Learning* 24 (1996). URL: https://doi.org/10.1007/BF00058655.

[16] Jason Brownlee. *8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset.* 2015. URL: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ (visited on 07/01/2022).

[17] Jason Brownlee. *Assessing and Comparing Classifier Performance with ROC Curves.* 2014. URL: https://machinelearningmastery.com/assessing-comparing-classifier-performance-roc-curves-2/ (visited on 07/01/2022).

[18] C. Bustos et al. “Explainable, automated urban interventions to improve pedestrian and vehicle safety”. In: *Transportation Research Part C: Emerging Technologies* 125 (2021), p. 103018. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103018. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21000498.

[19] Qing Cai et al. “Applying machine learning and google street view to explore effects of drivers’ visual environment on traffic safety”. In: *Transportation Research Part C: Emerging Technologies* 135 (2022), p. 103541. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103541. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21005234.

[20] Qing Cai et al. “Real-time crash prediction on expressways using deep generative models”. In: *Transportation Research Part C: Emerging Technologies* 117 (2020), p. 102697. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102697. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306124.

[21] Zhong Cao et al. “Trustworthy safety improvement for autonomous driving using reinforcement learning”. In: *Transportation Research Part C: Emerging Technologies* 138 (2022), p. 103656. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103656. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22000997.

[22] Marwa Chabbouh et al. “Multi-objective evolution of oblique decision trees for imbalanced data binary classification”. In: *Swarm and Evolutionary Computation* 49 (2019), pp. 1–22. ISSN: 2210-6502. DOI: https://doi.org/10.1016/j.swevo.2019.05.005. URL: https://www.sciencedirect.com/science/article/pii/S2210650218305054.

[23] N. V. Chawla et al. “SMOTE: Synthetic Minority Over-sampling Technique”. In: *Journal of Artificial Intelligence Research* 16 (June 2002), pp. 321–357. ISSN: 1076-9757. DOI: 10.1613/jair.953. URL: http://dx.doi.org/10.1613/jair.953.

[24] Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. “Editorial: Special Issue on Learning from Imbalanced Data Sets”. In: *SIGKDD Explor. Newsl.* 6.1 (June 2004), pp. 1–6. ISSN: 1931-0145. DOI: 10.1145/1007730.1007733. URL: https://doi.org/10.1145/1007730.1007733.

[25] NV Chawla et al. “SMOTE: Synthetic minority over-sampling technique.” In: *JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH* 16 (2002), pp. 321–357. ISSN: 10769757.

[26] Chao Chen and Mei-Ling Shyu. “Clustering-based binary-class classification for imbalanced data sets”. In: *2011 IEEE International Conference on Information Reuse Integration.* 2011, pp. 384–389. DOI: 10.1109/IRI.2011.6009578.

[27] Chuqiao Chen et al. “Spatial-temporal pricing for ride-sourcing platform with reinforcement learning”. In: *Transportation Research Part C: Emerging Technologies* 130 (2021), p. 103272. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103272. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21002849.

[28] Qijun Chen et al. "A deep neural network inverse solution to recover pre-crash impact data of car collisions". In: *Transportation Research Part C: Emerging Technologies* 126 (2021), p. 103009. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103009. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21000413.

[29] Tianyi Chen et al. "Predicting lane-changing risk level based on vehicles' space-series features: A pre-emptive learning approach". In: *Transportation Research Part C: Emerging Technologies* 116 (2020), p. 102646. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102646. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20305611.

[30] Zhiju Chen et al. "H-ConvLSTM-based bagging learning approach for ride-hailing demand prediction considering imbalance problems and sparse uncertainty". In: *Transportation Research Part C: Emerging Technologies* 140 (2022), p. 103709. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103709. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22001474.

[31] Andy H.F. Chow et al. "Adaptive signal control for bus service reliability with connected vehicle technology via reinforcement learning". In: *Transportation Research Part C: Emerging Technologies* 129 (2021), p. 103264. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103264. URL: https://www.sciencedirect.com/science/article/pii/S0968090X2100276X.

[32] Samantha J. Corrado et al. "A clustering-based quantitative analysis of the interdependent relationship between spatial and energy anomalies in ADS-B trajectory data". In: *Transportation Research Part C: Emerging Technologies* 131 (2021), p. 103331. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103331. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003351.

[33] Aimee E. Cox and Jessica B. Cicchino. "Continued trends in older driver crash involvement rates in the United States: Data through 2017–2018". In: *Journal of Safety Research* 77 (2021), pp. 288–295. ISSN: 0022-4375. DOI: https://doi.org/10.1016/j.jsr.2021.03.013. URL: https://www.sciencedirect.com/science/article/pii/S0022437521000463.

[34] Zhiyong Cui et al. "Learning traffic as a graph: A gated graph wavelet recurrent neural network for network-scale traffic prediction". In: *Transportation Research Part C: Emerging Technologies* 115 (2020), p. 102620. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102620. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19306448.

[35] Sina Dabiri et al. "A deep convolutional neural network based approach for vehicle classification using large-scale GPS trajectory data". In: *Transportation Research Part C: Emerging Technologies* 116 (2020), p. 102644. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102644. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20305593.

[36] Damien Dablain, Bartosz Krawczyk, and Nitesh V. Chawla. "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data." In: (2021).

[37] Xingyuan Dai et al. "DeepTrend 2.0: A light-weighted multi-scale traffic prediction model using detrending". In: *Transportation Research Part C: Emerging Technologies* 103 (2019), pp. 142–157. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.03.022. URL: https://www.sciencedirect.com/science/article/pii/S0968090X1830648X.

[38] Raj Deshmukh et al. "Temporal logic learning-based anomaly detection in metroplex terminal airspace operations". In: *Transportation Research Part C: Emerging Technologies* 126 (2021), p. 103036. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103036. URL: https://www.sciencedirect.com/science/article/pii/S0968090X2100067X.

[39] Imen Dhief et al. "A machine learned go-around prediction model using pilot-in-the-loop simulations". In: *Transportation Research Part C: Emerging Technologies* 140 (2022), p. 103704. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103704. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22001425.

[40] Loan N.N. Do et al. "An effective spatial-temporal attention based neural network for traffic flow prediction". In: *Transportation Research Part C: Emerging Technologies* 108 (2019), pp. 12–28. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.09.008. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19301330.

[41] Jiqian Dong et al. "Space-weighted information fusion using deep reinforcement learning: The context of tactical control of lane-changing autonomous vehicles and connectivity range assessment". In: *Transportation Research Part C: Emerging Technologies* 128 (2021), p. 103192. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103192. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21002084.

[42] Jan Drchal, Michal Čertický, and Michal Jakob. "Data-driven activity scheduler for agent-based mobility models". In: *Transportation Research Part C: Emerging Technologies* 98 (2019), pp. 370–390. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2018.12.002. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18306417.

[43] Wenbo Du et al. "Cooperative pursuit of unauthorized UAVs in urban airspace via Multi-agent reinforcement learning". In: *Transportation Research Part C: Emerging Technologies* 128 (2021), p. 103122. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103122. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21001418.

[44] Yuchuan Du et al. "Comfortable and energy-efficient speed control of autonomous vehicles on rough pavements using deep reinforcement learning". In: *Transportation Research Part C: Emerging Technologies* 134 (2022), p. 103489. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103489. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004757.

[45] Zouhair Elamrani Abou Elassad, Hajar Mousannif, and Hassan Al Moatassime. “A real-time crash prediction fusion framework: An imbalance-aware strategy for collision avoidance systems”. In: *Transportation Research Part C: Emerging Technologies* 118 (2020), p. 102708. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102708. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306239.

[46] Hamed Faroqi and Mahmoud Mesbah. “Inferring trip purpose by clustering sequences of smart card records”. In: *Transportation Research Part C: Emerging Technologies* 127 (2021), p. 103131. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103131. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21001509.

[47] Siyuan Feng et al. “Coordinating ride-sourcing and public transport services with a reinforcement learning approach”. In: *Transportation Research Part C: Emerging Technologies* 138 (2022), p. 103611. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103611. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22000572.

[48] Yoav Freund and Robert E Schapire. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”. In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. URL: https://www.sciencedirect.com/science/article/pii/S002200009791504X.

[49] Sergio Garrido et al. “Prediction of rare feature combinations in population synthesis: Application of deep generative modelling”. In: *Transportation Research Part C: Emerging Technologies* 120 (2020), p. 102787. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102787. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306975.

[50] Alexander Genser and Anastasios Kouvelas. “Dynamic optimal congestion pricing in multi-region urban networks by application of a Multi-Layer-Neural network”. In: *Transportation Research Part C: Emerging Technologies* 134 (2022), p. 103485. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103485. URL: https://www.sciencedirect.com/science/article/pii/S0968090X2100471X.

[51] Taha Ghasempour and Benjamin Heydecker. “Adaptive railway traffic control using approximate dynamic programming”. In: *Transportation Research Part C: Emerging Technologies* 113 (2020). ISTTT 23 TR_C-23rd International Symposium on Transportation and Traffic Theory (ISTTT 23), pp. 91–107. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18317285.

[52] Melita J. Giummarra et al. “A systematic review of the association between fault or blame-related attributions and procedures after transport injury and health and work-related outcomes”. In: *Accident Analysis & Prevention* 135 (2020), p. 105333. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2019.105333. URL: https://www.sciencedirect.com/science/article/pii/S0001457519303781.

[53] Hongren Gong et al. "Two-vehicle driver-injury severity: A multivariate random parameters logit approach". In: *Analytic Methods in Accident Research* 33 (2022), p. 100190. ISSN: 2213-6657. DOI: https://doi.org/10.1016/j.amar.2021.100190. URL: https://www.sciencedirect.com/science/article/pii/S2213665721000348.

[54] Ian J. Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. Ed. by Zoubin Ghahramani et al. 2014, pp. 2672–2680. URL: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

[55] Volkan Gumuskaya et al. "Integrating stochastic programs and decision trees in capacitated barge planning with uncertain container arrivals". In: *Transportation Research Part C: Emerging Technologies* 132 (2021), p. 103383. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103383. URL: https://www.sciencedirect.com/science/article/pii/S0968090X2100382X.

[56] Qiangqiang Guo et al. "Hybrid deep reinforcement learning based eco-driving for low-level connected and automated vehicles along signalized corridors". In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102980. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.102980. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21000164.

[57] Yi Guo and Jiaqi Ma. "DRL-TP3: A learning and control framework for signalized intersections with mixed connected automated traffic". In: *Transportation Research Part C: Emerging Technologies* 132 (2021), p. 103416. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103416. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004101.

[58] Mohammad Hesam Hafezi et al. "Ensemble learning activity scheduler for activity based travel demand models". In: *Transportation Research Part C: Emerging Technologies* 123 (2021), p. 102972. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.102972. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21000097.

[59] Yu Han et al. "A physics-informed reinforcement learning-based strategy for local and coordinated ramp metering". In: *Transportation Research Part C: Emerging Technologies* 137 (2022), p. 103584. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103584. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22000304.

[60] Henrick J. Haule et al. "Evaluating the effect of ramp metering on freeway safety using real-time traffic data". In: *Accident Analysis & Prevention* 157 (2021), p. 106181. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2021.106181. URL: https://www.sciencedirect.com/science/article/pii/S0001457521002128.

[61] G.C. Herbert. *Crash Report Sampling System: Imputation*. Tech. rep. DOT HS 812 795. National Highway Traffic Safety Administration, Sept. 2019.

[62] Ken Hidaka et al. "Generating pedestrian walking behavior considering detour and pause in the path under space-time constraints". In: *Transportation Research Part C: Emerging Technologies* 108 (2019), pp. 115–129. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.09.005. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18312713.

[63] Sebastián Hughes et al. "Evaluation of machine learning methodologies to predict stop delivery times from GPS data". In: *Transportation Research Part C: Emerging Technologies* 109 (2019), pp. 289–304. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.10.018. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18314645.

[64] Koichi Ito and Filip Biljecki. "Assessing bikeability with street view imagery and computer vision". In: *Transportation Research Part C: Emerging Technologies* 132 (2021), p. 103371. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103371. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003739.

[65] Urvashi Jaitley. *Comparing Different Classification Machine Learning Models for an Imbalanced Dataset*. 2019. URL: https://towardsdatascience.com/comparing-different-classification-machine-learning-models-for-an-imbalanced-dataset-fdae1af3677f (visited on 07/01/2022).

[66] Feifeng Jiang, Kwok Kit Richard Yuen, and Eric Wai Ming Lee. "A long short-term memory-based framework for crash detection on freeways with traffic data of different temporal resolutions". In: *Accident Analysis & Prevention* 141 (2020), p. 105520. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2020.105520. URL: https://www.sciencedirect.com/science/article/pii/S0001457519317713.

[67] Yan Jiao et al. "Real-world ride-hailing vehicle repositioning using deep reinforcement learning". In: *Transportation Research Part C: Emerging Technologies* 130 (2021), p. 103289. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103289. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003004.

[68] Arash Kalatian and Bilal Farooq. "Decoding pedestrian and automated vehicle interactions using immersive virtual reality and interpretable deep learning". In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102962. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102962. URL: https://www.sciencedirect.com/science/article/pii/S0968090X2030855X.

[69] Mark S. Kaplan et al. "The National Violent Death Reporting System: Use of the Restricted Access Database and Recommendations for the System's Improvement". In: *American Journal of Preventive Medicine* 53.1 (2017), pp. 130–133. ISSN: 0749-3797. DOI: https://doi.org/10.1016/j.amepre.2017.01.043. URL: https://www.sciencedirect.com/science/article/pii/S0749379717301101.

[70] Jintao Ke et al. “Predicting origin-destination ride-sourcing demand with a spatio-temporal encoder-decoder residual multi-graph convolutional network”. In: *Transportation Research Part C: Emerging Technologies* 122 (2021), p. 102858. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102858. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20307580.

[71] Waqar Ahmed Khan et al. “Hierarchical integrated machine learning model for predicting flight departure delays and duration in series”. In: *Transportation Research Part C: Emerging Technologies* 129 (2021), p. 103225. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103225. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21002394.

[72] Jinsoo Kim, Jae Hun Kim, and Gunwoo Lee. “GPS data-based mobility mode inference model using long-term recurrent convolutional networks”. In: *Transportation Research Part C: Emerging Technologies* 135 (2022), p. 103523. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103523. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21005052.

[73] Nishant Kumar and Martin Raubal. “Applications of deep learning in congestion detection, prediction and alleviation: A survey”. In: *Transportation Research Part C: Emerging Technologies* 133 (2021), p. 103432. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103432. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004241.

[74] Craig D. Lack, Kathryn S. Berkow, and Yuanxue Gao. “Insights into motor carrier crashes: A preliminary investigation of FMCSA inspection violations”. In: *Accident Analysis & Prevention* 155 (2021), p. 106105. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2021.106105. URL: https://www.sciencedirect.com/science/article/pii/S0001457521001366.

[75] Shir Meir Lador. *What metrics should be used for evaluating a model on an imbalanced data set? (precision + recall or ROC=TPR+FPR)*. 2017. URL: https://towardsdatascience.com/what-metrics-should-we-use-on-imbalanced-data-set-precision-recall-roc-e2e79252aeba (visited on 07/01/2022).

[76] Daniel A. Lazar et al. “Learning how to dynamically route autonomous vehicles on shared roads”. In: *Transportation Research Part C: Emerging Technologies* 130 (2021), p. 103258. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103258. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21002709.

[77] Namgil Lee, Heejung Yang, and Hojin Yoo. *A surrogate loss function for optimization of F_β score in binary classification with imbalanced data*. 2021. arXiv: 2104.01459 [cs.LG].

[78] Seunghyeon Lee et al. “An advanced deep learning approach to real-time estimation of lane-based queue lengths at a signalized junction”. In: *Transportation Research Part C: Emerging Technologies* 109 (2019), pp. 117–136. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j

.trc.2019.10.011. URL: https://www.sciencedirect.com/science/article/pii/S0968090X1830812X.

[79] Can Li et al. "Urban mobility analytics: A deep spatial–temporal product neural network for traveler attributes inference". In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102921. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102921. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20308202.

[80] Guofa Li et al. "Decision making of autonomous vehicles in lane change scenarios: Deep reinforcement learning approaches with risk awareness". In: *Transportation Research Part C: Emerging Technologies* 134 (2022), p. 103452. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103452. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004411.

[81] Guopeng Li, Victor L. Knoop, and Hans van Lint. "Multistep traffic forecasting by dynamic graph convolution: Interpretations of real-time spatial correlations". In: *Transportation Research Part C: Emerging Technologies* 128 (2021), p. 103185. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103185. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21002011.

[82] Jinjian Li et al. "Multi-models machine learning methods for traffic flow estimation from Floating Car Data". In: *Transportation Research Part C: Emerging Technologies* 132 (2021), p. 103389. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103389. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003867.

[83] Jinlong Li et al. "Domain adaptation from daytime to nighttime: A situation-sensitive vehicle detection and traffic flow parameter estimation framework". In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102946. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102946. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20308433.

[84] Junyi Li et al. "Transferability improvement in short-term traffic prediction using stacked LSTM network". In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102977. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.102977. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21000140.

[85] Xiaojiang Li et al. "A novel method for predicting and mapping the occurrence of sun glare using Google Street View". In: *Transportation Research Part C: Emerging Technologies* 106 (2019), pp. 132–144. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.07.013. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18311252.

[86] Zhenning Li et al. "Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning". In: *Transportation Research Part C: Emerging Technologies* 125 (2021), p. 103059. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103059. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21000851.

[87] Zhongcan Li et al. “Prediction of train arrival delays considering route conflicts at multi-line stations”. In: *Transportation Research Part C: Emerging Technologies* 138 (2022), p. 103606. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103606. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22000523.

[88] T. Lin et al. “Focal Loss for Dense Object Detection.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence, Pattern Analysis and Machine Intelligence, IEEE Transactions on, IEEE Trans. Pattern Anal. Mach. Intell* 42.2 (2020), pp. 318–327. ISSN: 0162-8828.

[89] Yan Liu et al. “Dynamic activity chain pattern estimation under mobility demand changes during COVID-19”. In: *Transportation Research Part C: Emerging Technologies* 131 (2021), p. 103361. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103361. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003636.

[90] Yang Liu, Zhiyuan Liu, and Ruo Jia. “DeepPF: A deep learning based architecture for metro passenger flow prediction”. In: *Transportation Research Part C: Emerging Technologies* 101 (2019), pp. 18–34. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.01.027. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18306806.

[91] Yi Liu et al. “A double standard model for allocating limited emergency medical service vehicle resources ensuring service reliability.” In: *Transportation Research Part C* 69 (2016), pp. 120–133. ISSN: 0968-090X.

[92] Fei Lu, Zichen Chen, and Huiyu Chen. “Lateral collision risk assessment of parallel routes in ocean area based on space-based ADS-B”. In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102970. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.102970. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21000085.

[93] Ying Lv et al. “Mobility pattern recognition based prediction for the subway station related bike-sharing trips”. In: *Transportation Research Part C: Emerging Technologies* 133 (2021), p. 103404. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103404. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004009.

[94] Lijing Ma and Shiru Qu. “A sequence to sequence learning based car-following model for multi-step predictions considering reaction delay”. In: *Transportation Research Part C: Emerging Technologies* 120 (2020), p. 102785. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102785. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306951.

[95] Tao Ma, Constantinos Antoniou, and Tomer Toledo. “Hybrid machine learning algorithm and statistical time series model for network-wide traffic forecast”. In: *Transportation Research Part C: Emerging Technologies* 111 (2020), pp. 352–372. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.12.022. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19303821.

[96] Wei Ma, Xidong Pi, and Sean Qian. “Estimating multi-class dynamic origin-destination demand through a forward-backward algorithm on computational graphs”. In: *Transportation Research Part C: Emerging Technologies* 119 (2020), p. 102747. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2020.102747`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X20306604`.

[97] Yang Ma et al. “A convolutional neural network method to improve efficiency and visualization in modeling driver’s visual field on roads using MLS data”. In: *Transportation Research Part C: Emerging Technologies* 106 (2019), pp. 317–344. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2019.07.018`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X19301536`.

[98] Nada Mahmoud et al. “Predicting cycle-level traffic movements at signalized intersections using machine learning models”. In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102930. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2020.102930`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X20308299`.

[99] Kunti Robiatul Mahmudah et al. “Classification of Imbalanced Data Represented as Binary Features”. In: *Applied Sciences* 11.17 (2021). ISSN: 2076-3417. DOI: `10.3390/app11177825`. URL: `https://www.mdpi.com/2076-3417/11/17/7825`.

[100] Eleni G. Mantouka, Panagiotis Fafoutellis, and Eleni I. Vlahogianni. “Deep survival analysis of searching for on-street parking in urban areas”. In: *Transportation Research Part C: Emerging Technologies* 128 (2021), p. 103173. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2021.103173`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X21001911`.

[101] Chao Mao, Yulin Liu, and Zuo-Jun (Max) Shen. “Dispatch of autonomous vehicles for taxi services: A deep reinforcement learning approach”. In: *Transportation Research Part C: Emerging Technologies* 115 (2020), p. 102626. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2020.102626`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X19312227`.

[102] Zhaobin Mo, Rongye Shi, and Xuan Di. “A physics-informed deep learning paradigm for car-following models”. In: *Transportation Research Part C: Emerging Technologies* 130 (2021), p. 103240. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2021.103240`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X21002539`.

[103] Amin Mohammadnazar, Ramin Arvin, and Asad J. Khattak. “Classifying travelers’ driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning”. In: *Transportation Research Part C: Emerging Technologies* 122 (2021), p. 102917. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2020.102917`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X20308160`.

[104] Alexandra S. Mueller and Jessica B. Cicchino. "Teen driver crashes potentially preventable by crash avoidance features and teen-driver-specific safety technologies". In: *Journal of Safety Research* 81 (2022), pp. 305–312. ISSN: 0022-4375. DOI: https://doi.org/10.1016/j.jsr.2022.03.007. URL: https://www.sciencedirect.com/science/article/pii/S0022437522000433.

[105] Matthias Müller-Hannemann et al. "Estimating the robustness of public transport schedules using machine learning". In: *Transportation Research Part C: Emerging Technologies* 137 (2022), p. 103566. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103566. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22000146.

[106] Rahul Nair and Anton Dekusar. "Keep it simple stupid! A non-parametric kernel regression approach to forecast travel speeds". In: *Transportation Research Part C: Emerging Technologies* 110 (2020), pp. 269–274. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.11.018. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19309350.

[107] Rahul Nair et al. "An ensemble prediction model for train delays". In: *Transportation Research Part C: Emerging Technologies* 104 (2019), pp. 196–209. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.04.026. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18317984.

[108] National Center for Statistics and Analysis. *Crash Report Sampling System analytical user's manual, 2016-2020.* Tech. rep. DOT HS 813 236. National Highway Traffic Safety Administration, Mar. 2022.

[109] Tu Dinh Nguyen et al. *Dual Discriminator Generative Adversarial Nets.* 2017. arXiv: 1709.03831 [cs.LG].

[110] NHTSA. *Crash Report Sampling System.* https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system. 2016-2020.

[111] Xavier Olive and Luis Basora. "Detection and identification of significant events in historical aircraft trajectory data". In: *Transportation Research Part C: Emerging Technologies* 119 (2020), p. 102737. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102737. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306513.

[112] Osama A. Osman et al. "A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data". In: *Accident Analysis & Prevention* 123 (2019), pp. 274–281. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2018.12.005. URL: https://www.sciencedirect.com/science/article/pii/S000145751831114X.

[113] Antonio Paez et al. "A systematic assessment of the use of opponent variables, data subsetting and hierarchical specification in two-party crash severity analysis". In: *Accident Analysis & Prevention* 144 (2020), p. 105666. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.a

ap.2020.105666. URL: https://www.sciencedirect.com/science/article/pii/S0001457520303298.

[114] Venktesh Pandey, Evana Wang, and Stephen D. Boyles. "Deep reinforcement learning algorithm for dynamic pricing of express lanes with multiple access locations". In: *Transportation Research Part C: Emerging Technologies* 119 (2020), p. 102715. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102715. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306306.

[115] Yanbo Pang et al. "Development of people mass movement simulation framework based on reinforcement learning". In: *Transportation Research Part C: Emerging Technologies* 117 (2020), p. 102706. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102706. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306215.

[116] Yutian Pang et al. "Data-driven trajectory prediction with weather uncertainties: A Bayesian deep learning approach". In: *Transportation Research Part C: Emerging Technologies* 130 (2021), p. 103326. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103326. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003314.

[117] Hyoshin Park, Deion Waddell, and Ali Haghani. "Online optimization with look-ahead for freeway emergency vehicle dispatching considering availability". In: *Transportation Research Part C: Emerging Technologies* 109 (2019), pp. 95–116. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.09.016. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19313129.

[118] Edwin P. D. Pednault, Barry K. Rosen, and Chidanand Apté. "Handling Imbalanced Data Sets in Insurance Risk Modeling". In: AAAI Workshop on Learning from Imbalanced Data Sets, 2000.

[119] Duc-Thinh Pham et al. "Deep reinforcement learning based path stretch vector resolution in dense traffic with uncertainties". In: *Transportation Research Part C: Emerging Technologies* 135 (2022), p. 103463. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103463. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004514.

[120] Xuewei Qi et al. "Deep reinforcement learning enabled self-learning control for energy efficient driving". In: *Transportation Research Part C: Emerging Technologies* 99 (2019), pp. 67–81. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2018.12.018. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18318862.

[121] Guoyang Qin et al. "Optimizing matching time intervals for ride-hailing services using reinforcement learning". In: *Transportation Research Part C: Emerging Technologies* 129 (2021), p. 103239. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103239. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21002527.

[122] T. Raghunathan et al. *IVEware: Imputation and Variation Estimation Software.* Version 0.3. 2016. URL: https://www.src.isr.umich.edu/software/iveware/.

[123] Trivellore E. Raghunathan et al. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models". In: *Survey Methodology* 27.1 (2001), pp. 85–95.

[124] Md Adilur Rahim and Hany M. Hassan. "A deep learning based traffic crash severity prediction framework". In: *Accident Analysis & Prevention* 154 (2021), p. 106090. ISSN: 0001-4575. DOI: `https://doi.org/10.1016/j.aap.2021.106090`. URL: `https://www.sciencedirect.com/science/article/pii/S0001457521001214`.

[125] Moynur Rahman, Min-Wook Kang, and Pranesh Biswas. "Predicting time-varying, speed-varying dilemma zones using machine learning and continuous vehicle tracking". In: *Transportation Research Part C: Emerging Technologies* 130 (2021), p. 103310. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2021.103310`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X21003181`.

[126] Judy Raj. *What to Do When Your Classification Data is Imbalanced.* 2019. URL: `https://towardsdatascience.com/what-to-do-when-your-classification-dataset-is-imbalanced-6af031b12a36` (visited on 07/01/2022).

[127] Felix Rempe, Philipp Franeck, and Klaus Bogenberger. "On the estimation of traffic speeds with Deep Convolutional Neural Networks given probe data". In: *Transportation Research Part C: Emerging Technologies* 134 (2022), p. 103448. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2021.103448`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X2100437X`.

[128] Baptiste Rocca. *Handling Imbalanced Datasets in Machine Learning.* 2019. URL: `https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28` (visited on 07/01/2022).

[129] Hector Rodriguez-Deniz, Mattias Villani, and Augusto Voltes-Dorta. "A multilayered block network model to forecast large dynamic transportation graphs: An application to US air transport". In: *Transportation Research Part C: Emerging Technologies* 137 (2022), p. 103556. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2022.103556`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X22000055`.

[130] Kamol Chandra Roy et al. "Predicting traffic demand during hurricane evacuation using Real-time data from transportation systems and social media". In: *Transportation Research Part C: Emerging Technologies* 131 (2021), p. 103339. ISSN: 0968-090X. DOI: `https://doi.org/10.1016/j.trc.2021.103339`. URL: `https://www.sciencedirect.com/science/article/pii/S0968090X21003429`.

[131] Matthias Schlögl. "A multivariate analysis of environmental effects on road accident occurrence using a balanced bagging approach". In: *Accident Analysis & Prevention* 136 (2020), p. 105398. ISSN: 0001-4575. DOI: `https://doi.org/10.1016/j.aap.2019.105398`. URL: `https://www.sciencedirect.com/science/article/pii/S0001457519308516`.

[132] Michael Schultz, Stefan Reitmann, and Sameer Alam. “Predictive classification and understanding of weather impact on airport performance through machine learning”. In: *Transportation Research Part C: Emerging Technologies* 131 (2021), p. 103119. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103119. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21001388.

[133] Georges Sfeir, Filipe Rodrigues, and Maya Abou-Zeid. “Gaussian process latent class choice models”. In: *Transportation Research Part C: Emerging Technologies* 136 (2022), p. 103552. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103552. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22000018.

[134] Amin Sharififar et al. “Addressing the issue of digital mapping of soil classes with imbalanced class observations.” In: *Geoderma* 350 (2019), pp. 84–92. ISSN: 0016-7061.

[135] Haotian Shi et al. “Connected automated vehicle cooperative control with a deep reinforcement learning approach in a mixed traffic environment”. In: *Transportation Research Part C: Emerging Technologies* 133 (2021), p. 103421. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103421. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004150.

[136] Qian Shi and Hui Zhang. “An improved learning-based LSTM approach for lane change intention prediction subject to imbalanced data”. In: *Transportation Research Part C: Emerging Technologies* 133 (2021), p. 103414. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103414. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004083.

[137] Zhenyu Shou and Xuan Di. “Reward design for driver repositioning using multi-agent reinforcement learning”. In: *Transportation Research Part C: Emerging Technologies* 119 (2020), p. 102738. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102738. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306525.

[138] Zhenyu Shou et al. “Multi-agent reinforcement learning for Markov routing games: A new modeling paradigm for dynamic traffic assignment”. In: *Transportation Research Part C: Emerging Technologies* 137 (2022), p. 103560. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103560. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22000092.

[139] Samira Soleimani et al. “A Comprehensive Railroad-Highway Grade Crossing Consolidation Model: A Machine Learning Approach”. In: *Accident Analysis & Prevention* 128 (2019), pp. 65–77. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2019.04.002. URL: https://www.sciencedirect.com/science/article/pii/S0001457518305736.

[140] Ali Soleymani. *Stop Using SMOTE to Treat Class Imbalance.* Apr. 2022. URL: https://towardsdatascience.com/stop-using-smote-to-treat-class-imbalance-take-this-intuitive-approach-instead-9cb822b8dc45 (visited on 06/29/2022).

[141] Pranjal Soni. *Handling Imbalanced Datasets with imblearn Library.* Oct. 2020. URL: https://medium.com/thecyphy/handling-imbalanced-datasets-with-imblearn-library-df5e58b968f4 (visited on 07/01/2022).

[142] Rebecca Spicer et al. "Frequency and cost of crashes, fatalities, and injuries involving disabled vehicles". In: *Accident Analysis & Prevention* 152 (2021), p. 105974. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2021.105974. URL: https://www.sciencedirect.com/science/article/pii/S0001457521000051.

[143] Patrick Stewart. *What data scientists keep missing about imbalanced datasets.* 2021. URL: https://medium.com/mlearning-ai/what-data-scientists-keep-missing-about-imbalanced-datasets-d1f10e808297 (visited on 07/01/2022).

[144] Rajesh Subramanian et al. *Transitioning to multiple imputation: a new method to impute missing blood alcohol concentration (BAC) values in FARS.* Tech. rep. National Center for Statistics and Analysis (US), 2002.

[145] Jie Sun and Jiwon Kim. "Joint prediction of next location and travel time from urban vehicle trajectories using long short-term memory neural networks". In: *Transportation Research Part C: Emerging Technologies* 128 (2021), p. 103114. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103114. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21001339.

[146] Ding-Wen Tan et al. "A feature selection model for binary classification of imbalanced data based on preference for target instances". In: *2012 4th Conference on Data Mining and Optimization (DMO).* 2012, pp. 35–42. DOI: 10.1109/DMO.2012.6329795.

[147] Jinjun Tang et al. "Multi-community passenger demand prediction at region level based on spatio-temporal graph convolutional network". In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102951. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102951. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20308482.

[148] Ruifan Tang et al. "A literature review of Artificial Intelligence applications in railway systems". In: *Transportation Research Part C: Emerging Technologies* 140 (2022), p. 103679. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103679. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22001206.

[149] Xindi Tang et al. "Online operations of automated electric taxi fleets: An advisor-student reinforcement learning framework". In: *Transportation Research Part C: Emerging Technologies* 121 (2020), p. 102844. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102844. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20307464.

[150] Ivan Tomek. "Two modifications of CNN". In: *IEEE transactions on Systems, Man and Communications, SMC* 6 (1976), pp. 769–772.

[151] Kazim Topuz and Dursun Delen. “A probabilistic Bayesian inference model to investigate injury severity in automobile crashes”. In: *Decision Support Systems* 150 (2021). Interpretable Data Science For Decision Making, p. 113557. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2021.113557. URL: https://www.sciencedirect.com/science/article/pii/S0167923621000671.

[152] Berkay Turan, Ramtin Pedarsani, and Mahnoosh Alizadeh. “Dynamic pricing and fleet management for electric autonomous mobility on demand systems”. In: *Transportation Research Part C: Emerging Technologies* 121 (2020), p. 102829. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102829. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20307336.

[153] Christian Eduardo Verdonk Gallego et al. “A machine learning approach to air traffic interdependency modelling and its application to trajectory prediction”. In: *Transportation Research Part C: Emerging Technologies* 107 (2019), pp. 356–386. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.08.015. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18316024.

[154] Leon Villavicencio et al. “Passenger Presence and the Relative Risk of Teen Driver Death”. In: *Journal of Adolescent Health* 70.5 (2022), pp. 757–762. ISSN: 1054-139X. DOI: https://doi.org/10.1016/j.jadohealth.2021.10.038. URL: https://www.sciencedirect.com/science/article/pii/S1054139X21005759.

[155] Chen Wang et al. “Adaptive ensemble of classifiers with regularization for imbalanced data classification”. In: *Information Fusion* 69 (2021), pp. 81–102. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2020.10.017. URL: https://www.sciencedirect.com/science/article/pii/S1566253520303869.

[156] Hong-Wei Wang et al. “Evaluation and prediction of transportation resilience under extreme weather events: A diffusion graph convolutional approach”. In: *Transportation Research Part C: Emerging Technologies* 115 (2020), p. 102619. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102619. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19305868.

[157] Jiawei Wang, Ruixiang Chen, and Zhaocheng He. “Traffic speed prediction for urban transportation network: A path based deep learning approach”. In: *Transportation Research Part C: Emerging Technologies* 100 (2019), pp. 372–385. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.02.002. URL: https://www.sciencedirect.com/science/article/pii/S0968090X1831043X.

[158] Jiawei Wang and Lijun Sun. “Dynamic holding control to avoid bus bunching: A multi-agent deep reinforcement learning framework”. In: *Transportation Research Part C: Emerging Technologies* 116 (2020), p. 102661. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102661. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20305763.

[159] Shenhao Wang, Qingyi Wang, and Jinhua Zhao. "Deep neural networks for choice analysis: Extracting complete economic information for interpretation". In: *Transportation Research Part C: Emerging Technologies* 118 (2020), p. 102701. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102701. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306161.

[160] Tong Wang, Jiahua Cao, and Azhar Hussain. "Adaptive Traffic Signal Control for large-scale scenario with Cooperative Group-based Multi-agent reinforcement learning". In: *Transportation Research Part C: Emerging Technologies* 125 (2021), p. 103046. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103046. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21000760.

[161] Wensi Wang et al. "A data-driven hybrid control framework to improve transit performance". In: *Transportation Research Part C: Emerging Technologies* 107 (2019), pp. 387–410. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.08.017. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18312567.

[162] Xinglei Wang et al. "Forecast network-wide traffic states for multiple steps ahead: A deep learning approach considering dynamic non-local spatial correlation and non-stationary temporal dependency". In: *Transportation Research Part C: Emerging Technologies* 119 (2020), p. 102763. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102763. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306756.

[163] Xingmin Wang et al. "Learning the max pressure control for urban traffic networks considering the phase switching loss". In: *Transportation Research Part C: Emerging Technologies* 140 (2022), p. 103670. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103670. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22001139.

[164] Xinwei Wang et al. "Aircraft taxi time prediction: Feature importance and their implications". In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102892. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102892. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20307920.

[165] Yibing Wang et al. "Ego-efficient lane changes of connected and automated vehicles with impacts on traffic flow". In: *Transportation Research Part C: Emerging Technologies* 138 (2022), p. 103478. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103478. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004642.

[166] Yuan Wang et al. "Enhancing transportation systems via deep learning: A survey". In: *Transportation Research Part C: Emerging Technologies* 99 (2019), pp. 144–163. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2018.12.004. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18304108.

[167] Marius Wegener et al. "Automated eco-driving in urban scenarios using deep reinforcement learning". In: *Transportation Research Part C: Emerging Technologies* 126 (2021), p. 102967.

ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.102967. URL: https://www.sciencedirect.com/science/article/pii/S0968090X2100005X.

[168] Jianan Wei et al. "New imbalanced bearing fault diagnosis method based on Sample-characteristic Oversampling TechniquE (SCOTE) and multi-class LS-SVM". In: *Applied Soft Computing* 101 (2021), p. 107043. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2020.107043. URL: https://www.sciencedirect.com/science/article/pii/S1568494620309819.

[169] Gary Weiss. "Learning to Predict Extremely Rare Events". In: *AAAI Workshop on Learning from Imbalanced Data Sets* (May 2000).

[170] Melvin Wong and Bilal Farooq. "A bi-partite generative model framework for analyzing and simulating large scale multiple discrete-continuous travel behaviour data". In: *Transportation Research Part C: Emerging Technologies* 110 (2020), pp. 247–268. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.11.022. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19300841.

[171] Melvin Wong and Bilal Farooq. "ResLogit: A residual neural network logit model for data-driven choice modelling". In: *Transportation Research Part C: Emerging Technologies* 126 (2021), p. 103050. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103050. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21000802.

[172] Yuankai Wu et al. "Differential variable speed limits control for freeway recurrent bottlenecks via deep actor-critic algorithm". In: *Transportation Research Part C: Emerging Technologies* 117 (2020), p. 102649. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102649. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20305647.

[173] Yuanyuan Wu, Haipeng Chen, and Feng Zhu. "DCL-AIM: Decentralized coordination learning of autonomous intersection management for connected and automated vehicles". In: *Transportation Research Part C: Emerging Technologies* 103 (2019), pp. 246–260. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.04.012. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18316279.

[174] Yang Xing et al. "Multi-scale driver behavior modeling based on deep spatial-temporal representation for intelligent vehicles". In: *Transportation Research Part C: Emerging Technologies* 130 (2021), p. 103288. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103288. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21002990.

[175] Meng Xu et al. "Designing van-based mobile battery swapping and rebalancing services for dockless ebike-sharing systems based on the dueling double deep Q-network". In: *Transportation Research Part C: Emerging Technologies* 138 (2022), p. 103620. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103620. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22000651.

[176] Hanyi Yang, Lili Du, and Jamshid Mohammadi. "A shock wave diagram based deep learning model for early alerting an upcoming public event". In: *Transportation Research Part C: Emerging Technologies* 122 (2021), p. 102862. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102862. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20307622.

[177] Jin-Ming Yang, Zhong-Ren Peng, and Lei Lin. "Real-time spatiotemporal prediction and imputation of traffic status based on LSTM and Graph Laplacian regularized matrix factorization". In: *Transportation Research Part C: Emerging Technologies* 129 (2021), p. 103228. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103228. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21002412.

[178] Shuguan Yang et al. "A deep learning approach to real-time parking occupancy prediction in transportation networks incorporating multiple spatio-temporal data sources". In: *Transportation Research Part C: Emerging Technologies* 107 (2019), pp. 248–265. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.08.010. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18313780.

[179] Junlin Yao and Ayman Moawad. "Vehicle energy consumption estimation using large scale simulations and machine learning methods". In: *Transportation Research Part C: Emerging Technologies* 101 (2019), pp. 276–296. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.02.012. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19302293.

[180] Ruoyu Yao et al. "A deep learning framework for modelling left-turning vehicle behaviour considering diagonal-crossing motorcycle conflicts at mixed-flow intersections". In: *Transportation Research Part C: Emerging Technologies* 132 (2021), p. 103415. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103415. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004095.

[181] Yingjun Ye, Xiaohui Zhang, and Jian Sun. "Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment". In: *Transportation Research Part C: Emerging Technologies* 107 (2019), pp. 155–170. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.08.011. URL: https://www.sciencedirect.com/science/article/pii/S0968090X19311301.

[182] Jinwon Yoon et al. "Transferable traffic signal control: Reinforcement learning with graph centric state representation". In: *Transportation Research Part C: Emerging Technologies* 130 (2021), p. 103321. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103321. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003272.

[183] Rongjie Yu et al. "Convolutional neural networks with refined loss functions for the real-time crash risk analysis". In: *Transportation Research Part C: Emerging Technologies* 119 (2020), p. 102740. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102740. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306549.

[184] Rongjie Yu et al. "Convolutional neural networks with refined loss functions for the real-time crash risk analysis." In: *Transportation Research Part C* 119 (2020). ISSN: 0968-090X.

[185] Xinlian Yu and Song Gao. "A batch reinforcement learning approach to vacant taxi routing". In: *Transportation Research Part C: Emerging Technologies* 139 (2022), p. 103640. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103640. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22000833.

[186] JUN-HAI ZHAI et al. "A Three-stage Method for Classification of Binary Imbalanced Big Data". In: *2020 International Conference on Machine Learning and Cybernetics (ICMLC)*. 2020, pp. 207–212. DOI: 10.1109/ICMLC51923.2020.9469568.

[187] Junhai Zhai, Jiaxing Qi, and Chu Shen. "Binary imbalanced data classification based on diversity oversampling by generative models". In: *Information Sciences* 585 (2022), pp. 313–343. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2021.11.058. URL: https://www.sciencedirect.com/science/article/pii/S0020025521011804.

[188] Junhai Zhai, Jiaxing Qi, and Sufang Zhang. "Binary Imbalanced Data Classification Based on Modified D2GAN Oversampling and Classifier Fusion". In: *IEEE Access* 8 (2020), pp. 169456–169469. DOI: 10.1109/ACCESS.2020.3023949.

[189] Jiawei Zhang et al. "A bi-level cooperative operation approach for AGV based automated valet parking". In: *Transportation Research Part C: Emerging Technologies* 128 (2021), p. 103140. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103140. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21001583.

[190] Jinlei Zhang et al. "Short-term origin-destination demand prediction in urban rail transit systems: A channel-wise attentive split-convolutional neural network method". In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102928. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102928. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20308275.

[191] Ke Zhang et al. "Multi-vehicle routing problems with soft time windows: A multi-agent reinforcement learning approach". In: *Transportation Research Part C: Emerging Technologies* 121 (2020), p. 102861. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102861. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20307610.

[192] Wei Zhang et al. "AdapGL: An adaptive graph learning algorithm for traffic prediction based on spatiotemporal neural networks". In: *Transportation Research Part C: Emerging Technologies* 139 (2022), p. 103659. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103659. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22001024.

[193] Wenwen Zhang et al. "Synthesizing neighborhood preferences for automated vehicles". In: *Transportation Research Part C: Emerging Technologies* 120 (2020), p. 102774. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102774. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20306847.

[194] Xiaohui Zhang et al. "Simultaneous modeling of car-following and lane-changing behaviors using deep learning". In: *Transportation Research Part C: Emerging Technologies* 104 (2019), pp. 287–304. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.05.021. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18308003.

[195] Xinyuan Zhang et al. "Online parking assignment in an environment of partially connected vehicles: A multi-agent deep reinforcement learning approach". In: *Transportation Research Part C: Emerging Technologies* 138 (2022), p. 103624. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103624. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22000699.

[196] Zhengchao Zhang et al. "A customized deep learning approach to integrate network-scale online traffic data imputation and prediction". In: *Transportation Research Part C: Emerging Technologies* 132 (2021), p. 103372. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103372. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21003740.

[197] Zhengchao Zhang et al. "Multistep speed prediction on traffic networks: A deep learning approach considering spatio-temporal dependencies". In: *Transportation Research Part C: Emerging Technologies* 105 (2019), pp. 297–322. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2019.05.039. URL: https://www.sciencedirect.com/science/article/pii/S0968090X18315389.

[198] Tingting Zhao and Yu Zhang. "Learning-based restoration sequence ordering for multi-site traffic signal failure". In: *Transportation Research Part C: Emerging Technologies* 135 (2022), p. 103522. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103522. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21005040.

[199] Ming Zheng et al. "An automatic sampling ratio detection method based on genetic algorithm for imbalanced data classification". In: *Knowledge-Based Systems* 216 (2021), p. 106800. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2021.106800. URL: https://www.sciencedirect.com/science/article/pii/S0950705121000630.

[200] Yunhan Zheng, Shenhao Wang, and Jinhua Zhao. "Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models". In: *Transportation Research Part C: Emerging Technologies* 132 (2021), p. 103410. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103410. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21004058.

[201] Dongqin Zhou and Vikash V. Gayah. "Model-free perimeter metering control for two-region urban networks using deep reinforcement learning". In: *Transportation Research Part C: Emerging Technologies* 124 (2021), p. 102949. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102949. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20308469.

[202] Meixin Zhu et al. “Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving”. In: *Transportation Research Part C: Emerging Technologies* 117 (2020), p. 102662. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102662. URL: https://www.sciencedirect.com/science/article/pii/S0968090X20305775.

[203] Xi Zhu et al. “An online updating method for time-varying preference learning”. In: *Transportation Research Part C: Emerging Technologies* 121 (2020), p. 102849. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2020.102849. URL: https://www.sciencedirect.com/science/article/pii/S0968090X2030749X.

[204] Xinting Zhu and Lishuai Li. “Flight time prediction for fuel loading decisions with a deep learning approach”. In: *Transportation Research Part C: Emerging Technologies* 128 (2021), p. 103179. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2021.103179. URL: https://www.sciencedirect.com/science/article/pii/S0968090X21001959.

[205] Apostolos Ziakopoulos and George Yannis. “A review of spatial approaches in road safety”. In: *Accident Analysis & Prevention* 135 (2020), p. 105323. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2019.105323. URL: https://www.sciencedirect.com/science/article/pii/S0001457519309893.

[206] Natalia Zuniga-Garcia et al. “Evaluation of e-scooters as transit last-mile solution”. In: *Transportation Research Part C: Emerging Technologies* 139 (2022), p. 103660. ISSN: 0968-090X. DOI: https://doi.org/10.1016/j.trc.2022.103660. URL: https://www.sciencedirect.com/science/article/pii/S0968090X22001036.