Dissertation Prospectus

Brad Burkman

Louisiana School for Math, Science, and the Arts

University of Louisiana at Lafayette

4 November 2022

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
- Feature Selection
- Cleaning Data
- Imbalanced Data
- Costs of Getting it Wrong

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
  - Accelerometer
  - Google Pixel
  - No Eyewitness
  - Send Ambulance?

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
  - No perfect database
  - Crash Report Sampling System
  - Louisiana data

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
  - Crash Report Sampling System
    - Public (scrubbed)
    - Non-representative sample of US
    - Some features imputed
  - Louisiana data
    - Not Public
    - Census
    - Raw

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
- Feature Selection

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
- Feature Selection
  - Time, Day, Weather
  - Location
  - Person?
  - Number of people?

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
- Feature Selection
- Cleaning Data

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
- Feature Selection
- Cleaning Data
  - CRSS: IVEware
  - Louisiana

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
- Feature Selection
- Cleaning Data
- Imbalanced Data

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
- Feature Selection
- Cleaning Data
- Imbalanced Data
  - Lots of tools, some relevant
  - Use all of them

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
- Feature Selection
- Cleaning Data
- Imbalanced Data
- Costs of Getting it Wrong

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
- Feature Selection
- Cleaning Data
- Imbalanced Data
- Costs of Getting it Wrong
  - Different costs
  - Class weights

Using two historical data sets, from the information available when police receive an automated crash report from a cell phone, build a model to recommend whether to automatically dispatch an ambulance.

- Cell Phone Automated Crash Reports
- Databases
- Feature Selection
- Cleaning Data
- Imbalanced Data
- Costs of Getting it Wrong

How I Got Here: Life

- U of Michigan
- Old Dominion U
- Wheaton College (Illinois)
- Dalian (Northeast China)
- SUNY Buffalo
- LSMSA
- UL
- Future

UNIVERSITY of LOUISIANA LAFAYETTE

How I Got Here: Computing

- Application Problem (2008)
  - Set Covering Problem
  - NP-Hard
- LSU Center for Computation and Technology
- SC and XSEDE Conferences
- Met People: Henry Neeman, Bob Panoff, Scott Lathrop, Kathy Traxler, Mark Jarrell, Juana Moreno, Box Leangsuksun
- LA-SiGMA RET (2010-2014)
- Sabbatical 2018-2019

How I Got Here: Dissertation Topic

- Algorithms and Reinforcement Learning with Dr. Jin (Spring & Fall 2019)
- Reinforcement Learning on the Rubik's Cube (December 2019)
- Louisiana Crash Report Data (February 2021)
- Application Problem
  - Dispatching ambulance from OnStar: May 2021
  - Cell phone: October 2021
- CRSS Data, Open Science (April 2022)

Dissertation Goals

Mastery of the Tools and Techniques of Research

Significant Contribution to the Field

Dissertation Goals

Mastery of the Tools and Techniques of Research

- Finding research question
- Literature review
- Finding appropriate datasets
- Cleaning and organizing data
- Handling imbalanced data
- Building models
- Interpreting results

Significant Contribution to the Field

Dissertation Goals

Mastery of the Tools and Techniques of Research

Significant Contribution to the Field

- New application question
- New dataset
- New imputation method for dataset
- New metrics
- New interpretation of class weights
- New combination of methods
- Open science

Dissertation Goals

Mastery of the Tools and Techniques of Research

Significant Contribution to the Field

Deep Dive: Order of Operations

- Binning and Imputing
- Which should we do first?
- Experiment
- Possible outcomes
- Interpreting results

Example: WEATHER

- CRSS Accident Data Set
  - 51 features, 20 of which we will use
  - 259,077 samples
  - 91,714 (35%) have some value missing
- WEATHER feature
  - 11 known values
  - 13,284 (5.1%) samples not known
    - Not Reported (12,636, 4.88%)
    - Reported as Unknown (648, 0.25%)

Weather: Binning by Correlation to Hospitalization

Value	and Meaning	% Samples	% Hospital	Bin
5	Fog, Smog, Smoke	0.35	21.70	0
3	Sleet or Hail	0.12	18.02	0
1	Clear	73.35	16.22	1
2	Rain	9.30	15.98	2
10	Cloudy	15.13	15.71	3
8	Other	0.06	15.18	4
6	Severe Crosswinds	0.06	14.18	4
12	Freezing Rain or Drizzle	0.03	13.61	4
11	Blowing Snow	0.05	12.58	4
4	Snow	1.54	12.36	4
7	Blowing Sand, Soil, Dirt	0.02	11.93	4

Imputing Missing Values

- Delete samples with missing values
- Assign most common value in feature
- Build a model using other features (IVEware)
  - Imputation and Variance Estimation Software
  - U Michigan Institute for Social Research
  - Sequential Regression Multivariate Imputation (SRMI)
  - Used by CRSS to impute *some* features

Experimental Method

1. CRSS ACCIDENT data set (259,077 samples)
2. In each feature, note proportion of missing values (in WEATHER, 13,284 samples $\div$ 259,077 $=$ 5.1%)
3. Drop all samples with missing data (167,363 left)
4. Store copy for ground truth
5. In each feature, erase a proportional number of values (167,363 $\times$ 5.1% $=$ 8,581)
6. Bin then Impute
7. Impute then Bin
8. Analyze crosstabs

Weather Crosstabs: Perfect Imputation

Ideal Imputation / Ground Truth	0	1	2	3	4
0	45	0	0	0	0
1	0	6198	0	0	0
2	0	0	818	0	0
3	0	0	0	1327	0
4	0	0	0	0	193

8581 missing values

8581 (100%) imputed correctly

Weather Crosstabs: Bin before Imputing

Bin - Impute Ground Truth	0	1	2	3	4
0	0	32	4	7	2
1	40	4518	550	962	128
2	9	569	81	140	19
3	4	959	111	216	37
4	0	137	14	40	2

8581 missing values
4817 (56.14%) imputed correctly
4818 (56.15 %) on second run

Weather Crosstabs: Impute before Binning

Impute - Bin Ground Truth	0	1	2	3	4
0	0	35	0	10	0
1	41	4555	556	912	134
2	6	600	58	135	19
3	10	978	118	204	17
4	2	143	15	30	3

8581 missing values
4820 (56.17%) imputed correctly
4776 (55.66%) on second run

Weather Crosstabs: Both Orders of Operation

Impute - Bin Bin - Impute	0	1	2	3	4
0	0	43	3	7	0
1	38	4601	542	917	117
2	8	546	71	121	14
3	10	989	115	215	36
4	3	132	16	31	6

8581 missing values

4893 (57.02%) imputed differently

Weather Crosstabs: Impute to Mode

Impute to Mode	0	1	2	3	4
Ground Truth					
0	0	45	0	0	0
1	0	6198	0	0	0
2	0	818		0	0
3	0	1327	0	0	0
4	0	193	0	0	0

8581 missing values
6198 (72.23%) imputed correctly

HOUR (binned) Correlation to HOSPITAL

Value and Meaning	Bin	% Samples	% Hospital
Late Night (23-4)	6	6.64	25.27
Evening (20-22)	5	9.71	20.13
Early Morning (5-6)	0	3.67	19.67
Early Evening (18-19)	4	12.00	16.13
Morning (7-10)	1	17.18	14.87
Mid Day (11-14)	2	24.18	14.76
Rush Hour (15-17)	3	26.36	13.83
Unknown	99	0.27	9.79

HOUR Crosstabs: Bin then Impute

Bin_Impute Ground_Truth	0	1	2	3	4	5	6
0	3	5	3	9	10	6	3
1	2	26	41	29	13	1	1
2	2	45	66	45	9	1	1
3	4	40	44	63	14	8	2
4	3	12	7	23	15	15	11
5	6	2	1	9	15	27	13
6	15	2	0	3	12	16	25

728 missing values
225 (30.91%) imputed correctly

Questions about Imputation

- In imputation, is “better than random” considered “good”?
- If the SRMI on a feature is really good, is the feature redundant?
- Is SRMI susceptible to the imbalanced data problem?
- If SRMI works well on one feature and mode imputation on another, should I mix and match the methods?

Challenges

- Imposter Syndrome
- Perfect the Enemy of the Good
- Hear, See, Do
- Making Time

Timeline: Paper Submission

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*

Timeline: Write Dissertation

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	Finish first draft of dissertation

Timeline: Write Dissertation

June 2023	Get feedback, Read papers, Rework, Write, and Revise
July 2023	Get feedback, Read papers, Rework, Write, and Revise
August 2023	Get feedback, Read papers, Rework, Write, and Revise
September 2023	Submit Dissertation

Timeline: Submit Dissertation

Mid Nov 2023	Deadline for Dissertation Defense *Preliminary Approval of Dissertation* form due
Late Nov 2023	Dissertation due on archival paper
15 Dec 2023	Graduation

Acknowledgements

- Dr. Miao Jin
- Malek Abujileh
- Committee Members

Questions?