

Highlights

Modeling the Need for an Ambulance based on Automated Crash Reports from Cell Phones

First Author, Second Author, Third Author, Fourth Author

- Supports transferability and benchmarking of different approaches on a public large-scale dataset. We have attached the code we used to perform the analysis on the Crash Report Sampling System.
- Novel Application motivated by Emerging Technology: Machine Learning Classification Models for Dispatching Ambulances based on Automated Crash Reports
- New Use of Dataset: Used Crash Report Sampling System (CRSS), which has imputed missing values for some features, but not all of the ones we wanted to use. For the first time we have seen, we used the software the CRSS authors use for multiple imputation (IVEware) to impute missing values in more features.
- Explicit Incorporation of Imbalanced Costs
- Explicit Incorporation of Political Dimensions
- Perennial Machine Learning Challenge: Imbalanced Datasets

Modeling the Need for an Ambulance based on Automated Crash Reports from Cell Phones

First Author^{a,b}, Second Author^a, Third Author^{a,c} and Fourth Author^c

^aSchool, University,

^bOther School,

^cOther Department, University,

ARTICLE INFO

Keywords:

Automated crash notification
Ambulance dispatch
Emergency medical services
Machine learning
Imbalanced Cost
Imbalanced Data
Imputation

ABSTRACT

New Google Pixel phones can automatically notify an emergency dispatcher if the phone detects the deceleration profile of a vehicular crash. Most crash notifications come from an eyewitness who can say whether an ambulance is needed, but the automated notification from the cell phone cannot provide that information directly. Should the dispatcher immediately send an ambulance before receiving an eyewitness report? There are three options: Always, Wait, and Sometimes. The “Always” option refers to sending an ambulance to every automatically reported crash, even though most of them will not be needed. In the “Wait” option, the dispatcher sends police, but always waits for a call from an eyewitness (perhaps the police) before sending an ambulance. In the “Sometimes” option, the dispatcher relies on a machine learning recommendation system to decide whether to immediately dispatch an ambulance, reserving the option to send one later based on an eyewitness report.

This paper explores one option for building a machine learning (ML) model for making a recommendation in the “Sometimes” option. Our goal is to build a model that returns, for each feature vector (crash report, sample), a value $p \in [0, 1]$ that increases with the probability that the person needs an ambulance. Then we choose a threshold θ such that we immediately send ambulances to those automated crash reports with $p > \theta$, and wait for eyewitness confirmation for those reports with $p < \theta$. In an actual implementation, the choice of θ is political, not technical, so we consider and interpret several options.

Once a threshold has been chosen, the costs of the false positives (FP) and false negatives (FN) in dispatching ambulances are very different. The cost of sending an ambulance when one is not needed (FP) is measured in dollars, but the cost of not promptly sending an ambulance when one is needed (FN) is measured in lives. Choosing such a tradeoff threshold is ethically problematic, but governments implicitly choose such a tradeoff when they set budgets for emergency services.

We consider and interpret several options for the decision threshold θ based on the political consideration, “How much will it cost?” How many automated ambulance dispatches are we willing to fund (FP + TP) for each one of them that’s actually needed (TP)? We will explore two versions of that question, the total and the marginal.

We show that the quality of the model depends highly on the input data available, and we considered three levels of data availability. The “Easy” level includes time of day and weather, data the emergency dispatcher has before the notification. The “Medium” level adds the age and sex of the cell phone user and information about the location. The “Hard” level adds information about the vehicle likely to be driven by the cell phone user and detailed and temporal information about the location, like lighting conditions and whether it is currently a work zone.


We used the data of the Crash Report Sampling System (CRSS) to validate our approach. We have applied new methods (for this dataset in the literature) to handle missing data, and we have investigated several methods for handling the data imbalance. To promote discussion and future research, we have included all of the code we used in our analysis.

1. Introduction

1.1. Outline

- Dataset

– CRSS

 FirstAuthor@gmail.com (F. Author)
ORCID(s):

- * 2016-2020, 2021
- * Over-represents more serious crashes
- Feature Selection and Engineering
- Discretization
- Imputing Missing Values
- Imbalanced Data
 - Can't use SMOTE
 - Class Weights
 - Focal Loss
 - Bagging and Boosting Methods
 - Moving the Discrimination Threshold
- Threshold Options
 - Choose the precision that is politically acceptable
 - Total Precision, including Prior Probability equals Posterior Probability
 - Marginal Precision
- Results and Conclusions

2. Total and Marginal Precision

Given a model and a choice of decision threshold θ , the total number of needed ambulances we send (TP) divided by the total number we send (FP + TP) is called the *precision*. Note that TP is all of the elements of the positive class with $p > \theta$, FP is all of the elements of the negative class with $p > \theta$, and FP+TP is all of the elements of either class with $p > \theta$.

The *marginal precision* at θ is the ratio of the number of positive samples to the total number of samples in the neighborhood of p around θ . The marginal precision the minimum probability that an ambulance sent is needed. In the language of economics, it is the probability that last ambulance sent is needed.

For example, if the decision makers are willing to send two unneeded ambulances (FP = $2k$ for some k) for every one that is needed (TP = $1k$), we look for the value of p where $\text{Prec} = \frac{1}{2+1} = 1/3$. If we want each ambulance sent to have at least a $1/3$ probability of being needed, then we look for the neighborhood of p where $m\text{Prec} = 1/3$.

The marginal precision is equivalent to the slope of the ROC curve, as there is an invertible mapping between them.

$$\begin{aligned}
 m\text{ROC} &= \frac{\Delta\text{TPR}}{\Delta\text{FPR}} = \frac{\Delta(\text{TP}/P)}{\Delta(\text{FP}/N)} \\
 &= \frac{(\Delta\text{TP})/P}{(\Delta\text{FP})/N} \quad (\text{because in a given model on a given data set, } P \text{ and } N \text{ are constant}) \\
 &= \frac{N}{P} \cdot \frac{\Delta\text{TP}}{\Delta\text{FP}} = \frac{N}{P} \cdot \frac{\Delta\text{TP}/\Delta p}{\Delta\text{FP}/\Delta p} = \frac{N}{P} \cdot \frac{\text{Pos}}{\text{Neg}} \\
 &= \frac{N}{P} \cdot \frac{1}{\frac{\text{Neg}}{\text{Pos}}} = \frac{N}{P} \cdot \frac{1}{\frac{\text{Neg}}{\text{Pos}} + 1 - 1} = \frac{N}{P} \cdot \frac{1}{\frac{\text{Neg}+\text{Pos}}{\text{Pos}} - 1} = \frac{N}{P} \cdot \frac{1}{\frac{1}{m\text{Prec}} - 1} = \frac{N}{P} \cdot \frac{m\text{Prec}}{1 - m\text{Prec}} \\
 m\text{Prec} &= \frac{P \cdot m\text{ROC}}{N + P \cdot m\text{ROC}}
 \end{aligned}$$

A challenge with calculating the marginal precision is choosing the margin ϵ for the neighborhood about p . If we make ϵ just large enough, the marginal precision will be a decreasing function of p and we will glean one value of θ

where $mPrec$ is closest to the goal. Because our data set is discrete, however, too small values of ϵ will yield some neighborhoods with few or no values of the positive or negative class. Because two of our model algorithms give most values of p rounded to two decimal places, we have chosen to use one hundred non-overlapping intervals of p ($\epsilon = 0.005$) for our analysis.

The table below gives the values for each of a hundred p neighborhoods for one of our models. Looking at $p = 0.45$ and 0.46 , for instance, for $p \in [0, 0.45]$ the model has correctly classified 117,225 of the 180,245 elements of the negative class and 26,573 of the 33,825 elements of the positive class. Moving from $p = 0.45$ to $p = 0.46$, the model correctly classifies 3,079 more elements of the negative class and 436 fewer elements of the positive class.

Claim: The precision is an increasing function is equivalent to

$$\frac{Pos}{Neg} < \frac{TP}{FP}$$

which is not necessarily true if we zoom in to a sufficiently small interval of p , because of the stochastic and discrete nature of our data set. Over sufficiently large intervals of p , however, it is generally true that precision is an increasing function of the decision boundary θ , being equivalent to the ROC curve curving down.

If the politicians have decided that they will trade off two immediately dispatched ambulances for each needed ambulance, then we choose the decision threshold θ at the value of p where $Prec = 0.33$, which happens around $p = 0.49$. Note that in this case we would be sending some ambulances to some crashes where there is only a 15% chance that the ambulance is needed. Similarly for other political decisions about total tradeoffs; we will investigate $Prec = 1/2$ and $Prec = 2/3$.

If the politicians decide that they want to automatically dispatch ambulances only to notifications where the likelihood that the victim requires an ambulance is greater than $1/3$, then choose the decision threshold θ at the value of p where $mPrec = 0.33$, which happens around $p = 0.68$. The marginal precision is much more volatile than the total precision, but we can narrow it down to somewhere in that region. At this value of θ over half, $13,617/(12,766 + 13,617) \approx 0.52$, of the ambulances that we automatically dispatch turn out to be needed. Similarly for other politically-chosen minimum percentages; we will also investigate $1/2$ and $2/3$.

Balanced Random Forest Classifier, Hard features, No Tomek undersampling, No class weights, Test set, Version 1

p	Neg	Pos	mPrec	TN	FP	FN	TP	Prec	Rec	\hat{p}
0.00	107	0	0.00	107	180,138	0	33,825	0.16	1.00	1.00
0.01	238	3	0.01	345	179,900	3	33,822	0.16	1.00	1.00
0.02	331	2	0.01	676	179,569	5	33,820	0.16	1.00	1.00
0.03	441	3	0.01	1,117	179,128	8	33,817	0.16	1.00	0.99
0.04	526	6	0.01	1,643	178,602	14	33,811	0.16	1.00	0.99
0.05	751	6	0.01	2,394	177,851	20	33,805	0.16	1.00	0.99
0.06	854	8	0.01	3,248	176,997	28	33,797	0.16	1.00	0.98
0.07	1,060	9	0.01	4,308	175,937	37	33,788	0.16	1.00	0.98
0.08	1,235	13	0.01	5,543	174,702	50	33,775	0.16	1.00	0.97
0.09	1,375	16	0.01	6,918	173,327	66	33,759	0.16	1.00	0.97
0.10	1,653	16	0.01	8,571	171,674	82	33,743	0.16	1.00	0.96
⋮										⋮
0.45	3,206	424	0.12	117,225	63,020	7,252	26,573	0.30	0.79	0.42
0.46	3,079	436	0.12	120,304	59,941	7,688	26,137	0.30	0.77	0.40
0.47	3,021	547	0.15	123,325	56,920	8,235	25,590	0.31	0.76	0.39
0.48	2,990	453	0.13	126,315	53,930	8,688	25,137	0.32	0.74	0.37
0.49	3,020	533	0.15	129,335	50,910	9,221	24,604	0.33	0.73	0.35
0.50	2,874	501	0.15	132,209	48,036	9,722	24,103	0.33	0.71	0.34
0.51	2,804	533	0.16	135,013	45,232	10,255	23,570	0.34	0.70	0.32
0.52	2,675	542	0.17	137,688	42,557	10,797	23,028	0.35	0.68	0.31
0.53	2,543	526	0.17	140,231	40,014	11,323	22,502	0.36	0.67	0.29
0.54	2,438	545	0.18	142,669	37,576	11,868	21,957	0.37	0.65	0.28
0.55	2,350	579	0.20	145,019	35,226	12,447	21,378	0.38	0.63	0.26
⋮										⋮
0.60	1,877	587	0.24	155,512	24,733	15,419	18,406	0.43	0.54	0.20
0.61	1,756	597	0.25	157,268	22,977	16,016	17,809	0.44	0.53	0.19
0.62	1,674	632	0.27	158,942	21,303	16,648	17,177	0.45	0.51	0.18
0.63	1,611	604	0.27	160,553	19,692	17,252	16,573	0.46	0.49	0.17
0.64	1,582	586	0.27	162,135	18,110	17,838	15,987	0.47	0.47	0.16
0.65	1,439	618	0.30	163,574	16,671	18,456	15,369	0.48	0.45	0.15
0.66	1,376	561	0.29	164,950	15,295	19,017	14,808	0.49	0.44	0.14
0.67	1,288	637	0.33	166,238	14,007	19,654	14,171	0.50	0.42	0.13
0.68	1,241	554	0.31	167,479	12,766	20,208	13,617	0.52	0.40	0.12
0.69	1,082	631	0.37	168,561	11,684	20,839	12,986	0.53	0.38	0.12
0.70	1,053	570	0.35	169,614	10,631	21,409	12,416	0.54	0.37	0.11
0.71	922	587	0.39	170,536	9,709	21,996	11,829	0.55	0.35	0.10
0.72	897	559	0.38	171,433	8,812	22,555	11,270	0.56	0.33	0.09
0.73	783	587	0.43	172,216	8,029	23,142	10,683	0.57	0.32	0.09
0.74	831	558	0.40	173,047	7,198	23,700	10,125	0.58	0.30	0.08
0.75	711	602	0.46	173,758	6,487	24,302	9,523	0.59	0.28	0.07
⋮										⋮
0.95	74	251	0.77	180,091	154	33,222	603	0.80	0.02	0.00
0.96	61	211	0.78	180,152	93	33,433	392	0.81	0.01	0.00
0.97	55	168	0.75	180,207	38	33,601	224	0.85	0.01	0.00
0.98	22	125	0.85	180,229	16	33,726	99	0.86	0.00	0.00
0.99	12	66	0.85	180,241	4	33,792	33	0.89	0.00	0.00
1.00	4	33	0.89	180,245	0	33,825	0	nan	0.00	0.00

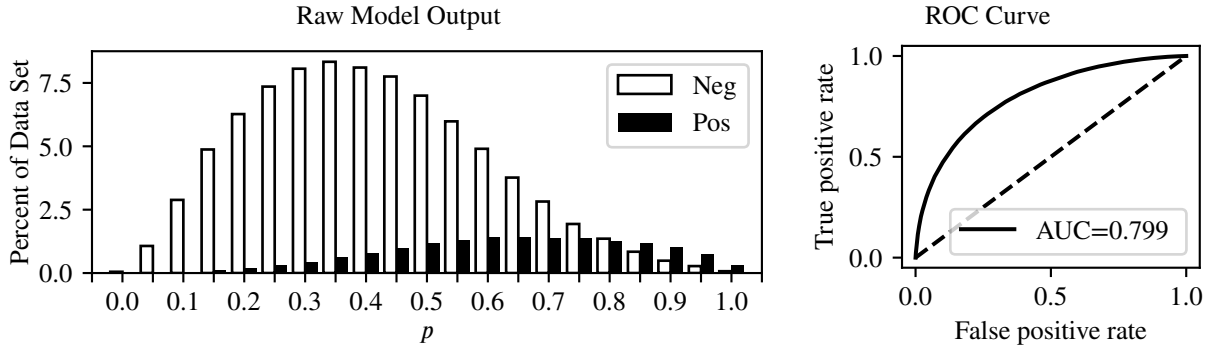
Visually on the histogram below, when we say that the precision of $1/3$ happens at $p \approx 0.50$, we mean that on the interval $p \in [0.50, 1.0]$, the area under the Neg curve is twice the area under the Pos curve. When we say that the marginal precision of $1/3$ happens at $p \approx 0.7$, we mean that the Neg bar at $p = 0.7$ is twice as tall as the Pos bar.

On the ROC curve below, $m\text{Prec} = 1/3$ happens at $p \approx 0.7$, on the curve at $(\text{FPR}, \text{TPR}) = (0.06, 0.37)$ because

$$\text{FPR} = \frac{\text{FP}}{\text{N}} = \frac{10,631}{180,091} = 0.06 \quad \text{and} \quad \text{TPR} = \frac{\text{TP}}{\text{P}} = \frac{12,416}{33,825} = 0.37$$

and at that point

$$m\text{ROC} = \frac{\text{N}}{\text{P}} \cdot \frac{m\text{Prec}}{1 - m\text{Prec}} = \frac{180,254}{33,825} \cdot \frac{1/3}{1 - 1/3} \approx 3$$



3. Literature Review

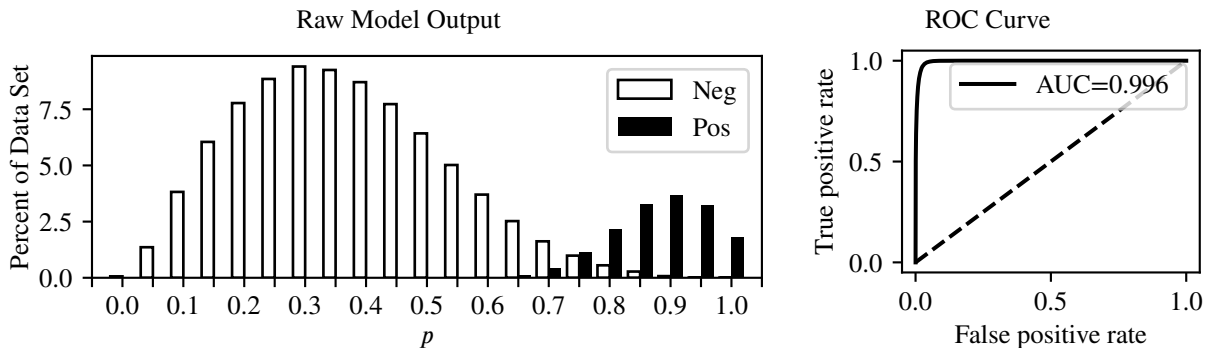
4. Dataset

5. Methods

5.1. Analysis of Results

Our ML algorithms assign to each sample (feature vector, crash person) a probability $p \in [0, 1]$ that the person needs an ambulance. The histogram below left shows the percentage of the dataset in each range of p , showing the percentages for the negative class (“Does not need an ambulance”) and the positive class (“Needs an ambulance”). On the right, the Receiver Operating Characteristic (ROC) curve, and particularly the area under the curve (AUC), is a metric for how well the model separates the two classes, with $AUC = 1.0$ being perfect and $AUC = 0.5$ (the dashed line) being just random assignment with no insight.

We would love to have results like in the graphs below, where the machine learning (ML) algorithm nearly perfectly separates the two classes. There is some overlap between $p = 0.6$ and $p = 0.8$ with some samples the algorithm misclassifies, but the model clearly separates most samples. Having an AUC of 0.996 would be amazing.



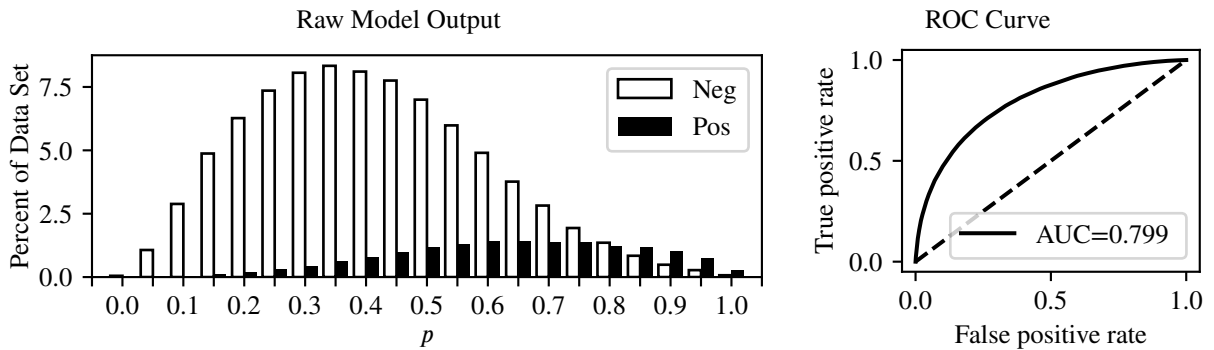
Unfortunately, our test results do not look quite that nice. They do not separate the two classes as well. Some distributions are clustered to one side or in the middle. Some models give the results in $p \in [0, 1]$ rounded to two decimal places so that we cannot hope for a level of detail beyond that, and one algorithm, Bagging, gives p rounded to only one decimal place.

Let us look at some examples. In all of them, AUC is in the range $[0.7, 0.8]$, so the various models separate the positive and negative classes about equally well overall, with none being dramatically better or worse. We will later show how we investigated which models do a better job in the ranges of interest.

BRFC_Hard_Tomek_0_alpha_0_5_v1_Test

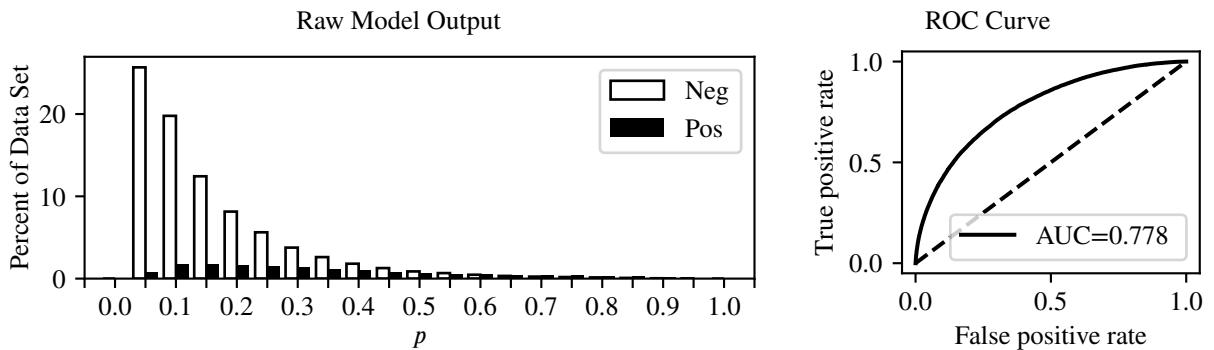
This model does not separate the negative and positive classes as well as the ideal, giving a much lower AUC (area under the ROC curve). These results are actually from the same model as the ideal above, but the ideal are the results on the training set and below on the test set, showing overfitting.

In these results, the 100 most frequent values comprised 93% of the results, meaning that, while there is some noise making the distribution look continuous, it is mostly discrete to two decimal places, so we cannot hope for fine detail in tuning the decision threshold.



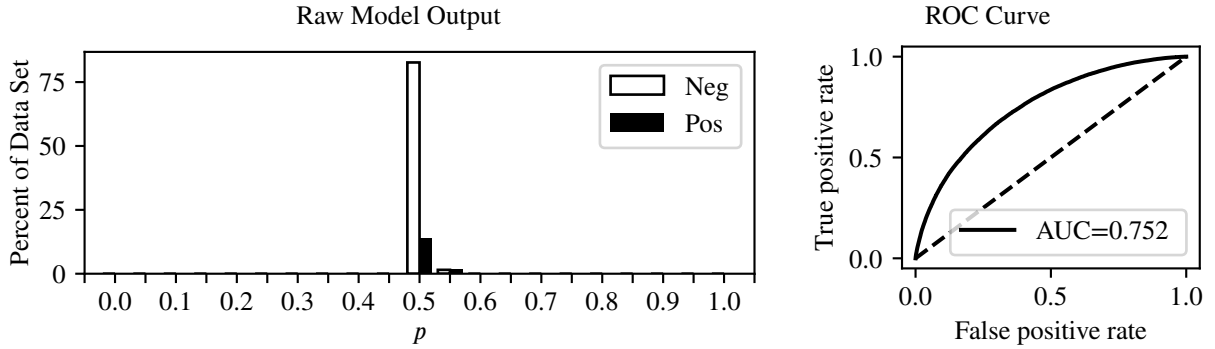
KBFC_Hard_Tomek_0_alpha_0_5_gamma_0_0_v1_Test

This model is almost as effective at separating the two classes (ROC = 0.778), but the distribution is skewed to the left. Its results were nearly continuous, with the 214,070 samples returning 210,157 unique values of p , so we can fine tune the decision threshold.

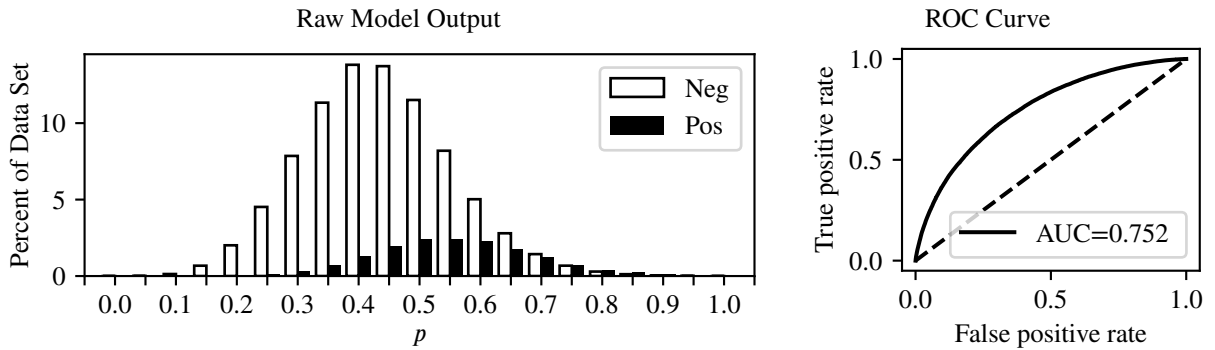


AdaBoost_Hard_Tomek_0_v1_Test

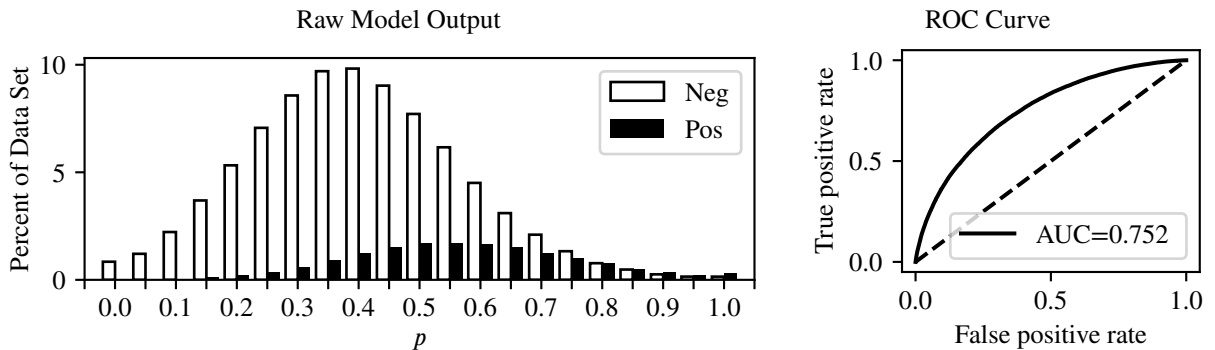
In this model the values are clustered very tightly, but in that small range the 214,070 samples return 210,442 different values of p , so there is much diversity that we can't see in this representation.



To make a useful visualization of the results where we can see the interplay between the negative and positive classes, we can transform the data. A transformation that preserves rank will have no effect on the ROC curve. [Cite] For the graph below, we mapped the smallest value in the set to 0 and the largest to 1.

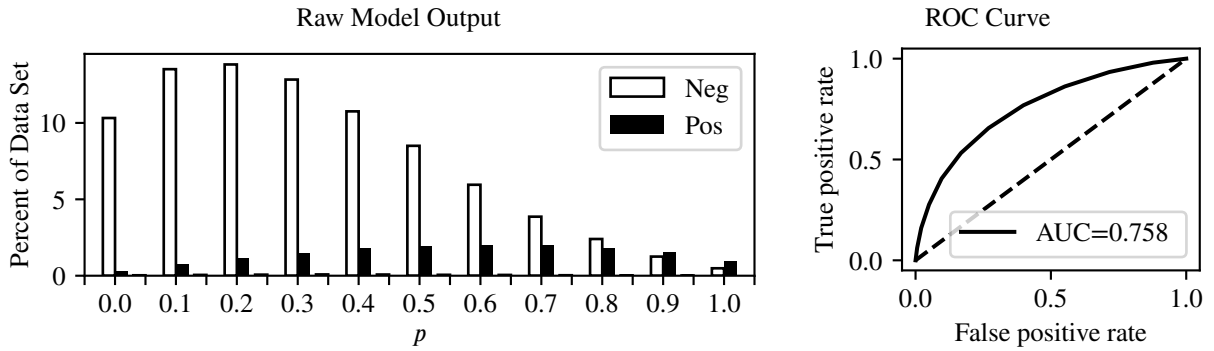


The distribution has long tails, so we can make a more useful visualization by truncating the ends. For this graph we mapped the 0.01 quantile to 0 and the 0.99 quantile to 1 leaving the center 98% of the distribution and truncated the ends. Our goal in clipping the tails is to make all of the models' results have approximately the same granularity when we choose the decision thresholds that give us the (politically) desired results.



Bagging_Hard_Tomek_0_v1_Test

This model returned 217 different values, but most of them were rare. Taking out the 5% of the data set with the least frequent values, 95% of the samples had only 10 values of p . It may be a useful model, but we will not be able to fine tune the decision threshold.



Other stuff

6. Results

7. Conclusions

8. Discussion

9. Future Work

10. To Do, Notes to Self

Funding Statement

Conflict of Interest

The authors have no relevant financial or non-financial interests to disclose.

Acknowledgements

[STUDENT] contributed to this work in the [FUNDED PROGRAM]

Data Availability

The CRSS data is publicly available at

<https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system>

11.

CRedit authorship contribution statement

First Author: Conceptualization, Investigation, Writing - original draft, Visualization. **Second Author:** Supervision, Methodology, Writing - review and editing. **Third Author:** Investigation, Methodology. **Fourth Author:** Data curation, Writing - review and editing.

References