ON THE NATURE OF NEGATIVE SAMPLING: HOW NON-ACCIDENT DATA HELPS US UNDERSTAND ACCIDENT OCCURRENCE

Peter Way
Department of Computer Science and Engineering
Center for Urban Informatics and Progress (CUIP)
University of Tennessee at Chattanooga
Chattanooga,TN 37403
peter-d-way@mocs.utc.edu

Jeremiah Roland
Department of Computer Science and Engineering
Center for Urban Informatics and Progress (CUIP)
University of Tennessee at Chattanooga
Chattanooga,TN 37403
jeremiah-roland@mocs.utc.edu

Mina Sartipi
Department of Computer Science and Engineering
Center for Urban Informatics and Progress (CUIP)
University of Tennessee at Chattanooga
Chattanooga,TN 37403
mina-sartipi@utc.edu

Word Count: 4734 words + 6 table(s) $\times$ 250 = 6234 words

Submission Date: August 1, 2019

ABSTRACT

In the process of studying event case data, such as traffic accident occurrences, it can often be difficult to gain a solid understanding of the meaning of a data record given there is often no opposing data to learn from. An example of this would be traffic data: There can be a plethora of data regarding when accidents occurred, but it becomes troublesome to find "non-accident" data as it technically does not exist. Negative sampling is a method of creating negative examples from a set of positive examples of a dataset. The usage of negative sampling is traditionally found in language based or numerical research questions. This paper outlines and analyzes the process of creating a variety of different negative sampling techniques, through various negative sample generation methods, each of which providing different results. The best performing results yielded a 92% accuracy in accident prediction and a 94% accuracy in non-accident prediction.

Keywords: Negative Sampling, Accident Prediction, Machine Learning

INTRODUCTION

Negative sampling is a method of creating negative examples from the existing collection of positive examples of a dataset. This technique of data generation is primarily only explored in natural language processing (NLP) or numerical research environments. However, negative sample creation and usages are now being sought after in the research of traffic patterns and accidents, as well as other smart city applications. It is critical to understand the various available types of negative sampling techniques, and which of these types may be best applied to answer a given research question. The positive samples explored here are traffic accident records from Hamilton County, Tennessee, and include temporal and spatial specifics from the accident location, as well as weather and roadway specifications. Various negative sampling techniques are explored, most of which are temporal and spatial reliant. These are specific aspects that numerical and language based negative techniques have not addressed previously, and which directly impact traffic based samples.

RELATED WORKS

A case study was conducted by (*1*) on predicting traffic accidents by utilizing and comparing the results of four different classification models of prediction. In this study, a method of generating non-accident data was performed and called negative sampling. For each positive example (accident), the value of only one feature was changed among hour, day, and road ID, the resulting sample was then checked for a positive (match found) or negative (no match found) result. Once all negative sampling methods were conducted, the team concluded the study with triple the number of negatives than positives, roughly a 75/25 split of data.

The team of (*2*) performed similar tests with accident prediction and negative sampling. Antoine et al. created their negatives through a process akin to brute force. Time and location information of the accidents in their dataset were examined and every single possible combination of them was generated, keeping only 0.1% of these newly created negatives. This method resulted in 2.3 million negatives for their dataset.

Additional related work discussed for negative sample creation may not be specifically traffic related, but their concepts may be well applied for such purposes. As an example, in (*3*) four strategies of negative sampling (local sampling, distance sampling, uniform sampling, and refined sampling) were studied for language processing applications. These four strategies were applied in exploration of Yahoo! question and answer community forums. *Local Sampling* negatives are those close to the existing positive sample by some given measure of approximation. This measure is able to be linguistically handled, or based on the actual vector's space. *Distance Sampling* negatives are those as distinct and different from the positive entries as capability allows. This ensures the data is correctly clustered in the given space of study. *Uniform Sampling*, simply said, is the random selection of negatives within the given space. This ensures that the entire space to be explored is represented equally, without preference to similarities or lack there of. *Refined Sampling* was defined as the combination of Local and Distance styled sampling, with the pursuit of a model capable of spanning clustered embeddings within a single category, as well as different categories. (*3*) also outlined some rules for negative samples; negative samples should be i) as similar as possible to positive samples to increase the model's discriminative abilities, ii) as different as possible to positive examples to avoid feeding the model conflicting information, and iii) representative of the entire space of negative samples.

This paper is partially supported by the NSF US Ignite 1647161.

Another three unique divisions of negative sample creation are presented by (*4*) within the negative sampling realm. They are presented as incompatible relations, domain specific rules, and random samples. *Incompatible Relations* are relations that always, or almost always, conflict with the relation wished to be extracted (*4*). In the case of previous traffic accident prediction project, an incompatible relation would be between generation of negative samples that exactly match current positive samples. If a generated negative sample has a certain time, date, and location, then positive samples cannot exist with the same time, date, and location, as there cannot be a non-accident where an accident was recorded. *Domain Specific Rules* are negative samples that are highly specific towards the particular data one is exploring (*4*). Similar to the above mentioned example, one cannot have a non-accident with the same time and location parameters. *Random Samples* deal with marking some current data as negative evidence.

RESEARCH METHODOLOGY

Data

All variables used in the negative sampling procedures and the creation of the given data set are shown in Table 1, along with a brief explanation of each variable. Grid Blocks, one of the variables used throughout this paper, refers to the image spaces seen in Figure 1. Each block seen is a Grid Block covering a 0.2 square mile area. Take note of the orientation of the grid block layout, as originally no orientation factor was chosen for the grid layout. Upon closer inspection, it was noted that there were many streets in Hamilton County that were being bisected into multiple grid blocks as the orientation of the roadways did not match up with the orientation of the grid blocks. Due to this, it was decided to match the grid block layout's orientation with Hamilton County's roadway layout to have better coverage of the roadway network. Furthermore, with the grid blocks matching the alignment of the roadway network, one may more easily visualize and analyze traffic trends throughout the city and the specific roadways.

TABLE 1 Data Features Used in Study

Variable	Description
Accident	Binary variable for accident occurrence
Hour	Hour of entry
UnixTime	Unix timestamp representation of entry
DayFrame	Time frame of day entry occurred
WeekDay / WeekEnd	Binary variables representing weekend/weekday
Clear/Cloudy/Rain/Fog/Snow	Binary variables (1 - present, 0 - not present)
RainBefore	Binary variable (1 - present, 0 - not present)
GridBlock	Numeric ID of GridBlock of entry
Grid Col / Grid Row	Column and row within grid of entry's GridBlock
Highway	Binary variable (1 - present, 0 - not present)
Land Use Mode	Type of surrounding area (Ex. Commercial, Urban, etc)
Road Count	Count of roadways within GridBlock of entry

Machine Learning Model

All tests performed within this work were done so with the machine learning model Multilayer Perceptron (MLP). The central reason behind choosing an MLP model for the given machine

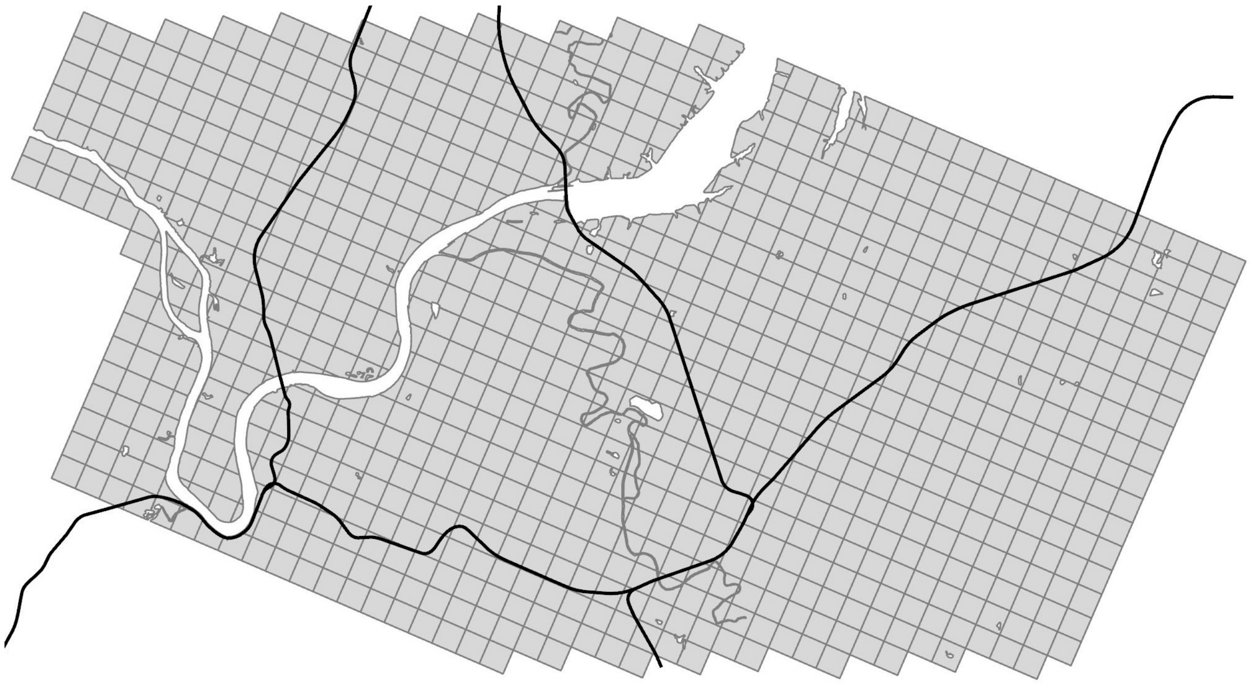

FIGURE 1 Grid Layout of Hamilton County used in Testing. Note Grid Block orientation aligns with local roadway network. Bolded black lines represent major interstates/highways. White segments in image convey bodies of water, whose Grid Blocks are ignored in model creation/testing.

learning technique is that the model itself best suits the project's data. MLPs are very flexible with the use of data, which is extremely beneficial to this project as the given dataset is very complex and intricate. Inputs are also labelled for classification prediction, which MLPs are suitable for. The details of the architecture used by the model are displayed in Table 2. Compilation was provided by MSE (mean squared error) with Nadam as the architecture optimizer. This particular testing was originally completed in this team's previous work (*5*).

Furthermore, research into the various models used for accident prediction has shown that different regression style models examine traffic flow differently, and as such, lead to varying results (*6*). An example of this previous research shows that Poisson distribution proved valuable in accident frequency analysis relating to accident frequency modeling. Poisson also prevailed over traditional linear regression in highway safety applications (*7*). Additionally, Negative Binomial models are useful in exploration of crash severity, as shown in previous works (*8*). Ordered logit/probit models are commonly applied, although usage of these highly depends on the levels of injury severity (*8*).Within previous binary level injury severity studies, many research teams chose to apply binary logistic modeling (*9–11*). To close, (*12*) applied ordered regression modelling to investigate five injury levels which ranged from no injury to fatal.

Table 2 displays the basic structural layout of the MLP model. Note in the Node column, a specific numerical value is that provided for the number of nodes used. For the different tests performed for this project (see Table 4), it was decided to have a method in place where instead of manually adjusting how many variables would be used for the three layers, a simple subtraction equation was put in place to set the number of nodes per layer based on the number of variables

TABLE 2 MLP Neural Network Architecture

Layer	Location	Type	Node	Activation
1	Input	Dense	# of Variables	Sigmoid
2	Hidden	Dense	# of Variables - 5	Sigmoid
3	Hidden	Dropout	-	Sigmoid
4	Hidden	Dense	# of Variables - 10	Sigmoid

supplied to the model. Note that this method requires there to be no less than 10 variables present for the model to use.

Creating Negative Samples

Despite the plethora of data available for analysis, it proved difficult to discover meaningful prediction results from them. This was due to the dataset consisting solely of positive examples for accident occurrences, as this interfered with any attempts at finding the important features in the process of accident occurrence and prediction. The results of (*1*) introduced the idea of implementing a negative sampling procedure for generating non-accident records. The procedure involves changing a single value of an accident record (hour, date, location) and checking if there is a matching accident record for the newly altered record. For example, if an accident occurred in hour 4, a new random hour was chosen between 0 and 23, excluding hour 4 for that day (*1*). The newly altered record was compared to all other accident records in the dataset to find any possible match. If no match was found, then the newly altered record was saved as a negative sample (non-accident). This process was repeated for every single accident entry in the dataset, and was done for each of the other two variable entries (date and location). This resulted in an increase in their dataset containing roughly 3 times more negative samples than positive samples. This team's process of negative sample generation was somewhat followed, which provided a similar increase in total data in our dataset.

After completing the procedure above, issues arose with accurate accident forecasting. That is, using the machine learning model to actually predict where accidents will occur in a given day. Due to this, it was decided to take a different approach to negative sample (NS) generation. Instead of changing only a single value of an accident record, a more varied approach in non-accident generation was used. This varied approach involved changing all of the given spatial and temporal variables (time, date, location) for a single accident record and finding any matching records. This process was repeated 9 times for each accident record in an attempt to reach a 90/10 split in data (90 percent non-accidents, and 10 percent accidents). The concept for a greater number of non-accidents came from an article written by (*13*) that discussed the importance of having a greater amount of negative examples of an event class scenario when the positive examples of the specific event are rare by nature. Given the inherent rarity of accidents occurring, the premise of maintaining the rarity of the accident's occurrence holds true in this project's circumstance as well, thus the particular 90/10 method. The third method of generating negative samples is very similar to the second method just described, but instead keeping the grid block (location) variable the same. This change in methodology was to find if the changing of location played a significant role in the quality of negative samples produced.

Negative Ratios

The ratio of negative to positive samples required for a scenario greatly depends on the given research question. (*14*) performed an examination of traffic accidents in Utah, exploring how it is important to have enough negative samples to clearly convey the rare occurrences of accidents, but not so many as to create a severe class imbalance. As mentioned, severe class imbalance leads to heavy bias toward the higher count occurrence. Conversely, training a model with an even split of non-accident and accident data may instruct the model that accidents and non-accidents occur with the same level of frequency. Now that the idea of varying ratios of negative to positive data has been introduced, the varying splits utilized by the aforementioned data may be explored further.

Original Modeling Split 66-33 The negatives created at this stage of research were greatly inspired by (*1*), and included shifting the Hour or Date variable to a new position independently. That is, if the accident occurred on January first, at nine in the morning, two different negatives would be created.

Increased Negative Sampling Split 75-25 This split was built upon the original modeling split, as location negative samples were added to the dataset. For these, the route ID and roadway segment were changed to another route ID and roadway segment combination.

Even Split 50-50 The even split was built upon the increased negative sampling split. It was believed that there was a detrimental class imbalance between positive and negative samples in the dataset, so work began to even out the split between positives and negatives. To accomplish this, the negative samples were scanned and every 3rd negative sample was retained, effectively cutting each negative sample case in thirds while retaining the original span of the negatives created.

'Rare' Circumstance Split > 90-10 This type of split is seen in several of the following methods used for negative sample generation in this paper. This type of dataset was used to see how much of an impact an overwhelming amount of negative samples would have on model performance and accident prediction,while retaining the 'rarity' of accident occurrence.

Sampling Types

The **Temporal Shift** explored here involves shifting either one or both of the temporal variables Hour and Weekday, while freezing the Grid Block variable's value. For example, an accident occurring at 6pm on Monday in Grid Block 32 could possibly have a negative at 6pm on Saturday in Grid Block 32. To retrieve negatives in this method, a list of accident records was compiled into a tabular format, where the column represented the hour (0 to 23) and the row represented the grid block. If a cell in the table featured a 0 value, it would mean an accident was not recorded at that hour in that grid block. This method of negative sampling was compounded upon further by the creation of negative samples for each weekend day that passed between the beginning of the year and up to the day that the data stopped. For testing, this negative sampling was performed on part of the 2019 dataset, which covered from January 1st 2019 to May 22nd 2019. Therefore, if one were creating negatives for weekends, then an additional 50 negative samples would be created for each instance of a negative sample (as there were 50 weekend days between beginning of the year and May 22). The reasoning for using this smaller sampling of data was to validate this specific type of negative sampling, as it takes a significant amount of time for the negatives to be created.

An extension of the temporal shift described above, the **Grid Fix** version of negative sampling also freezes Grid Block when creating negative samples. However, for this method of negative sample generation each record in the accident list was examined and the hour and date of the record were changed. After this was completed, the remaining accidents were queried for any matching records with the altered date and time. This process was performed 9 times for every entry, resulting in up to 9 additional negatives per positive. This edition of negative sampling produced the least number of negative samples due to its restrictive nature.

The **Spatial Shift** edition of this project's negative sampling involves shifting the Grid Block of the accident entry to create a negative sample. Both time sensitive aggregated variables were frozen, meaning that the new entry would occur at the same time of day, and would retain the same 'WeekDay' or 'WeekEnd' designation as the accident entry, while having its Grid Block changed. To retrieve negatives in this method, the steps listed abovewere repeated in the Temporal Shift description. However, for this testing negatives were created for each of the weekdays that had passed between the beginning of the year and the end of the current 2019 data, which was 126 business days.

The **Total Shift** edition of negative sampling exploration went about the selection of negatives in a different way than the spatial and temporal methods. For each positive sample available, the hour, date, and grid block of the entry were changed and the dataset queried for any matches for the newly created entry. This process was completed 9 times for each entry, providing up to 9 additional negatives per positive. This type of negative sampling is referred to as "Random" negative sampling, due to its inherent nature of changing all three of the variables previously discussed. This type of negative sampling is most similar to the negative sampling technique used in (*2*).

TABLE 3 DayFrame Time Coverage

DayFrame	Hours Covered
DayFrame 1	0 - 4 and 19 - 23 (Overnight)
DayFrame 2	5 - 9 (Morning rush)
DayFrame 3	10 - 13 (Lunch hours)
DayFrame 4	14 - 18 (Evening rush)

When conducting tests using these different negative sampling techniques, the terms "cut" and "full" are used in regard to negative samples. Full refers to the entire set of negative samples created through the respective method, while cut refers to a trimmed version of the negatives. This trimmed version was obtained based on aggregated temporal information, namely DayFrame and Weekday/Weekend. For example, if a method of negative sampling produced 2 negatives, each negative's Hour variable was aggregated into the DayFrame variable, which values represent certain hour intervals of the day. See Table 3 for hour breakdown of each DayFrame. Figure 2 illustrates the distribution of accident hotspots across the four DayFrames, highlighting the high intensity of accidents within the GridBlocks where the highway/interstates of the area can be found. Once properly aggregated, if the two created negatives have the same DayFrame, Weekday value, and Grid Block, then one of the negatives are dropped so only 1 negative entry with that specific DayFrame, Weekday, and Grid Block remains. This was done to better represent the raw data as well as simplify the model's input variables.

When conducting different tests on the previously discussed methods of negative sample generation, 6 different variable combinations were used. Table 4 displays the different variables dropped for each test, as well as the chronological order the tests were performed in. The reasoning behind particular variables being dropped for each test was to isolate the effect of each variable in terms of the model's performance. For example, the Clear variable was dropped since its value represents the absence of the other aggregated weather variables (Rain, Fog, Snow, Cloudy).

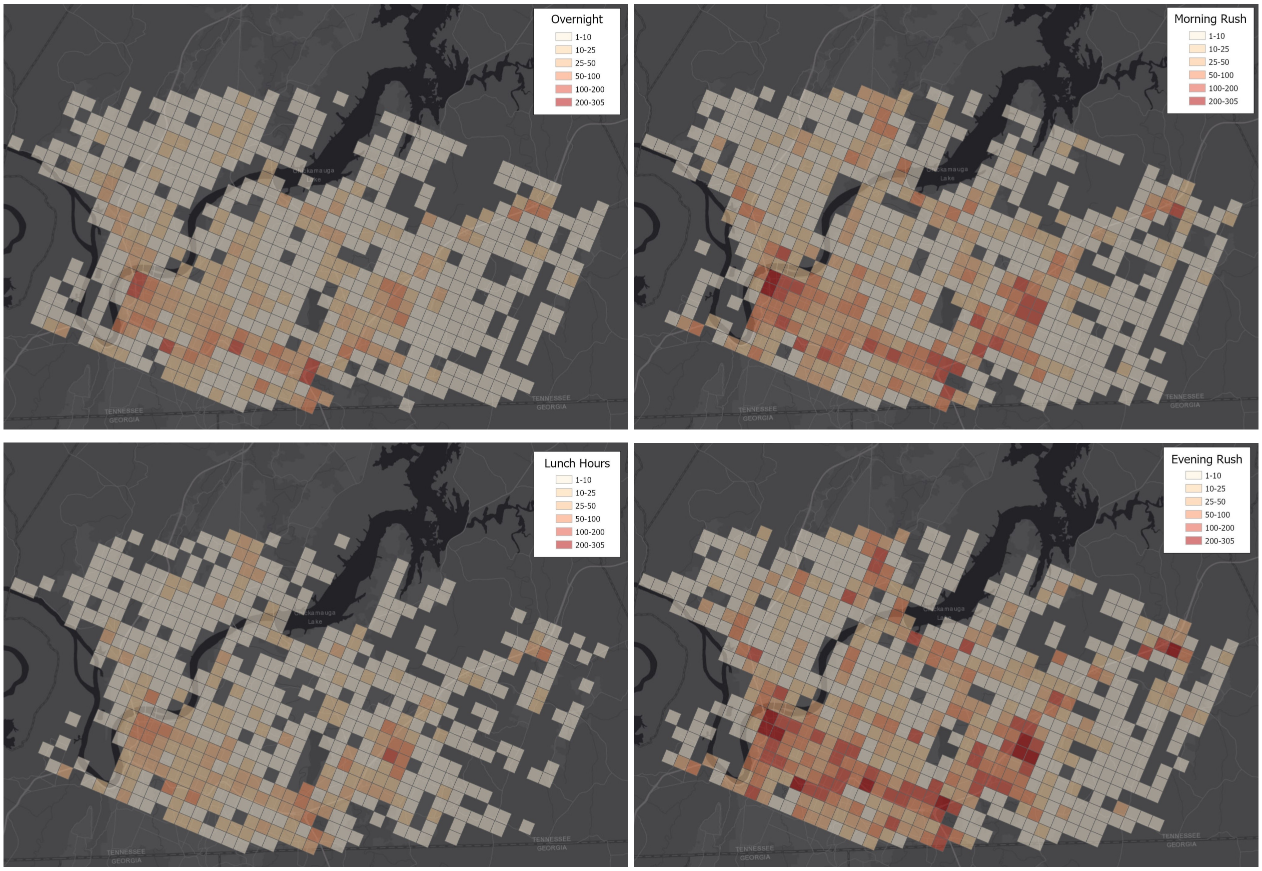

FIGURE 2 Accident Hotspots by DayFrame. DayFrames are in order from left to right (DayFrame 1, 2, 3, 4). The complete explanations of these DayFrames can be found in Table 3. Note: High accident hotspots correspond to Grid Blocks with highways within throughout all DayFrames (Placement of which can be examined in Figure 1).

RESULTS

For the **Original Modeling Split**, an Area Under the Curve (AUC) score of 81.37% was reported. Across the 3000 epochs used for testing the accuracy of the training set was marginally higher than the testing set, with a final accuracy of 77.50% on training and 76.92% on testing. The loss finished at 0.1537 for training, while the testing loss was marginally higher, at 0.1602 on the last epoch.

For the remaining model testing results in this paper, the performance of the model will be based upon the percent correct 1 and correct 0 scores. These two values represent the amount of correct positive and negative samples the model was able to predict. For example, if the testing dataset had 25% accident data (represented as 1 in accident), and 75% non-accident data (represented as 0 in accident), a Percent Correct 1 score of 0.816 means the model was able to predict 81.6% of accidents in the dataset. The reasoning behind this decision is that by using the % correct values, a more concrete understanding of the model's performance can be gleaned. By looking at a model's accuracy value, one is not able to derive the necessary information from it to tell whether the model can be trusted. For example, in Table 5 the Spatial Shift training and testing accuracies were both 99% yet the % correct 1 value was a lowly 37.5%, while the % correct 0 value was 100.

TABLE 4 Data Features Used in Tests

Test	Variables Dropped
Test 1	None
Test 2	Hour, WeekEnd, GridBlock, Clear
Test 3	DayFrame, GridBlock, Unix
Test 4	DayFrame, GridBlock, Hour
Test 5	Hour, Unix, GridBlock
Test 6	DayFrame, Unix

From this, it can be gathered that the Spatial Shift model was predicting a high amount of non-accidents. One would not have been able to understand this critical information by merely looking at the training and testing accuracies alone. Attention was also paid to the correlation matrices produced by these different models (represented as TN, FP, FN, and TP). The complete report of these initial results is shown by Table 5, as well as Figure 3.

The **Temporal Shift** model performed best upon the first test, completing with accuracies of 96.194% and 96.116% respectively for training and testing. The loss of both training and testing were some of the lowest found in all model tests, at 0.032 for both training and testing alike. The false positive rate of this model was rather good, with only 627 records, or roughly .7% of the total. The positive results continued with a 99.2% correct prediction score for the negative records. However, the model suffered with predicting accident records correctly, with only 50.4% of accident records being correctly identified. This model's dataset was roughly 93.5% negative data, as the negative/positive ratio was largely decided by the data itself based on the aforementioned methodology. This meant there were 5439 accident records, and 80170 negatives in the testing set.

The **Spatial Shift** model overall reported a very high training and testing accuracy across all test versions. As such, the loss for both training and testing remained very low. The false positive rate of this model was excellent as well, with 5 records. This positive outlook continued with the best test run from this model, with 100% of negative records being properly identified. However, once again the model performed abysmally with predicting positive records. Unfortunately, only 37.5% of accident records were correctly labeled. This version of the dataset also featured the most extreme ratio of negatives to positives, with 99.5% negative, and only .5% positive records. Once again, this cut was largely due to the data itself, with its high number of grid blocks versus the smaller amount of temporal variables. This version of negative sampling featured just 2,680 positive records and 490,882 negative records for a complete count of 493,562 entries in the testing set.

Full Grid Fix testing also reported high training and testing accuracies, with the best test option being test five. This particular run reported 94.6% on both training and testing accuracy. Loss also was mirrored on training and testing, at 0.046. The false positive rate of this model reported rather well, at only 1% of all records being reported in this category. This version of negative sampling also reported a high number of negative predictions, at 98.8% correct. Once again, however, only 43.4% of accidents were correctly predicted. This negative sampling featured a slightly lower bias toward negatives than the previously two mentioned, with 92.5% negatives, and 7.5% positive. To be precise, the Full Grid Fix style negative sampling had 13,250 positive entries, and 162,656 negative entries for a total of 175,906 records in the testing set.

The **Cut Grid Fix** testing performed best on test three, with 95.13% training and 94.834%

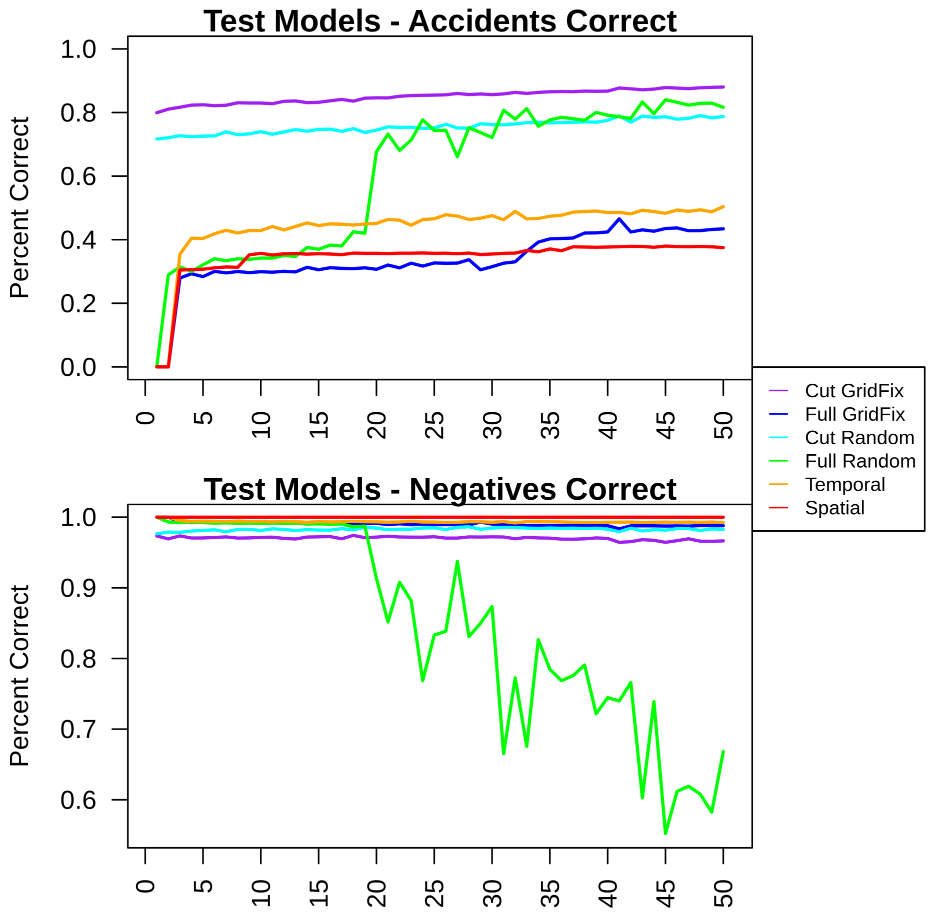

FIGURE 3 Percentages of Positive and Negative Entries correctly predicted via Various Models over the 50 training cycles. Strange performance of the Full Random model is thought to be due to possible conflicts between negative and positive record overlap once data aggregation is completed.

testing accuracy. Loss of both training and testing was reported as 0.043. This test performed best out of the best 6 of all negative sampling versions in regards to AUC, with 96.7%. This positive trend continued through with 90.5% of accidents and 96.5% of negatives correctly labeled. False positives were quite low for the number of entries, at a count of 1181 or 2.5% of all records. False negatives were similarly low, at 1252, or 2.7% of all records. This type of negative sampling had 13,201 positive entries, and 33,894 negative entries for a total testing set count of 47.095 entries. This leads to the dataset being roughly 73% negative and 28% positive, rather close to the generally accepted 75-25 split.

For the **Full Random** testing, test 6 provided the best results with a training and testing accuracy of 67.96 and 68.01, respectively. The standout for this dataset was an overall high number of false positives (FP). This outcome was expected due to the overwhelming amount of negative samples to positive samples in this particular dataset. There were a total of 164,367 entries with 151,139 negative samples and 13,228 positive samples, falling close to the 90/10 split. Due to the negative samples being the favored learning trait in this dataset, it is fair to say that the model would be more akin to predicting non-accidents (negatives) than accidents (positives).

For the **Cut Random**, test 1 gave the best results with training and testing accuracy being 94.54 and 94.39, respectively. The accuracies of this test have taken a significant boost from the full random testing, with an additional increase in the Percent Correct 0 predictions, which are up from 66.8% in the full random test to 98.3%. However, the Percent Correct 1 predictions slightly dropped down to 78.7% from 81.6% in the full random test. One of the best improvements of this version of the model came in the form of a significantly lower number of false positives, dropping from 50,147 in full random test down to 917. This is likely due to the better balance between the number of positives and negatives in this dataset; the cut random dataset contained a total of 66,351 entries with 53,148 negative samples and 13,203 positive samples, falling close to an 80/20 dataset split.

TABLE 5 Best Performing Test Runs for Negative Sample Datasets

NS Type	Train Acc	Train Loss	Test Acc	Test Loss	AUC	TN	FP	FN	TP	% Correct 1	% Correct 0
Cut GridFix	95.13	0.043	94.83	0.043	0.967	32713	1181	1252	11949	90.5	96.5
Full GridFix	94.62	0.046	94.62	0.046	0.84	160690	1966	7499	5751	43.4	98.8
Cut Random	94.55	0.046	94.39	0.046	0.957	52231	917	2807	10396	78.7	98.3
Full Random	67.97	0.04	68.02	0.214	0.84	100992	50147	2428	10800	81.6	66.8
Temporal Shift	96.19	0.032	96.12	0.032	0.92	79543	627	2698	2741	50.4	99.2
Spatial Shift	99.65	0.003	99.66	0.003	0.789	490877	5	1675	1005	37.5	100

TABLE 6 Ratio Test Runs for Best Performing Negative Sample Datasets.

NS Type	Train Acc	Train Loss	Test Acc	Test Loss	AUC	TN	FP	FN	TP	% Correct 1	% Correct 0
Cut GridFix (75/25)	94.81	0.046	94.84	0.044	0.966	33092	714	1718	11571	87.1	97.9
Cut GridFix (50/50)	93.75	0.056	93.24	0.057	0.966	10644	654	1007	12250	92.4	94.2
Full GridFix (75/25)	85.13	0.115	84.96	0.114	0.863	39463	1171	6939	6367	47.9	97.1
Full GridFix (50/50)	75.08	0.171	75.19	0.167	0.831	11173	2443	4217	9003	68.1	82.1
Cut Randoms (75/25)	94.89	0.044	94.74	0.044	0.957	52023	1035	2458	10835	81.5	98.0
Cut Randoms (50/50)	89.05	0.083	88.62	0.083	0.952	12021	1209	1813	11508	86.4	90.9
Full Randoms (75/25)	85.48	0.107	85.44	0.106	0.896	35523	2305	5129	8098	61.2	93.9
Full Randoms (50/50)	81.58	0.135	81.37	0.132	0.893	11344	2528	2507	10640	80.9	81.8
Temporal Shift (75/25)	86.86	0.096	87.05	0.093	0.914	15328	817	1961	3347	63.1	94.9
Temporal Shift (50/50)	83.12	0.128	83.53	0.122	0.905	4439	873	899	4550	83.5	83.6
Spatial Shift (75/25)	87.68	0.091	87.56	0.093	0.901	7930	246	1108	1604	59.1	97.0
Spatial Shift (50/50)	82.11	0.130	81.09	0.130	0.893	2219	503	522	2175	80.6	81.5

Ratio Tests

When exploring classification data problems, many experts will recommend the use of specific ratios of negative samples to positive entries. However, there are many differing opinions on what exactly those ratios should be. For the sake of simplicity, two different yet commonly accepted ratios of data were selected to be explored here. These two are 75% negative, 25% positive, with the second ratio being an equal 50% division between positive and negative data. All of the aforementioned types of negative sampling were retained, but were restricted to just enough entries to roughly fulfill the previously mentioned ratios. The full results of the different split tests can be seen in Table 6.

All of the above mentioned testing presented the even split producing the highest percent of correctly predicted positive entries, as demonstrated in Figure 4. This phenomenon can be seen in Figure 4. As with the previous testing, the Cut version of the Grid Fix negatives performed the best overall in predicting accidents, completing the 50 cycle training with roughly 92.4% of positive entries correctly predicted. The second and third best performers (Full GridFix and Cut Random) had very similar performance, ending with 87.1% and 86.4% respectively. As for the prediction of negative entries, the 75-25 split data sets outperformed the 50-50 split, with the vast majority of the 50-50 tests falling below the 85% mark.

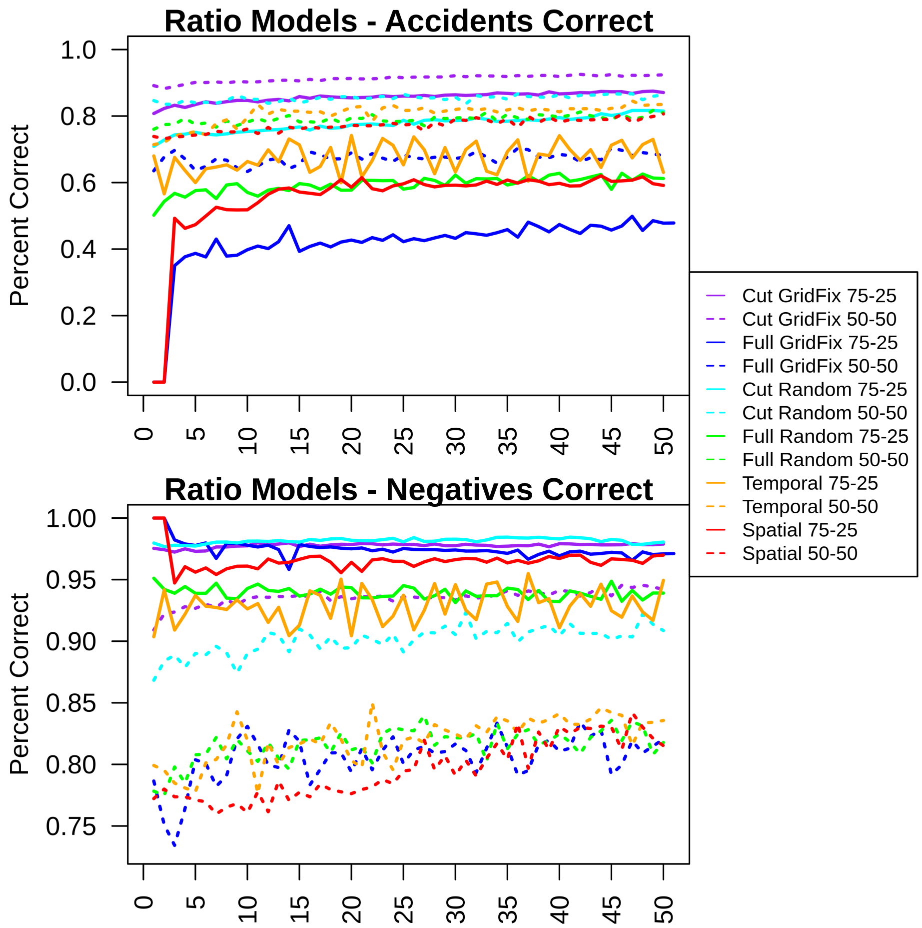

FIGURE 4 Percentages of Positive and Negative Entries correctly predicted via Various Ratio Models over the 50 training cycles

CONCLUSIONS

As mentioned, negative sampling had only previously explored primarily in a natural language processing based or numerical research environment. However, the creation and usage of negative sampling is now being sought after in the research of traffic patterns, accidents, and various smart city research questions. This paper explored many different negative sampling techniques, many of which take into account both temporal and spatial concerns that previous research into negative sampling had not addressed. It was found that for the purposes of accident prediction, fixing the Grid Block parameter and altering the Hour and WeekDay variables produced the best result in predicting traffic accident records, with 92.4% of traffic accidents being correctly labeled, and 94.2% of negative samples correctly labeled. Thus, it can be stated for this application and data, that a temporal shift with an even split between negatives and positives is the most accurate route to correctly predict traffic accident records. This is quite contrary to the original hypothesis regarding the relative rarity of accidents in daily occurrences, where hundreds of vehicles may pass a given area at a given time without incident. Therefore it must be reiterated that the specific negative sampling technique and ratio of data must be deciphered for each unique research situation.

ACKNOWLEDGEMENTS

We would like to thank the City of Chattanooga, Hamilton County Emergency Communications District, Tennessee Department of Transportation, Chattanooga Department of Transportation, and DarkSky for supplying data for this research. Furthermore, we would like to acknowledge the contributions to the paper and related project by Dr. Eric LaFlamme from Plymouth University for his help in understanding the principles behind analyzing complex data and negative sampling techniques, as well as Connor Firat of the Interdisciplinary Geospatial Technology Lab (IGTLab) at the University of Tennessee at Chattanooga for his help in creating the visualizations of Hamilton county and spatial attributes used in the project.

REFERENCES

1. Yuan, Z., X. Zhou, Y. Yang, J. Tamerius, and R. Mantialla, Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study. *Proceedings of 6th International Workshop on Urban Computing*, 2017.
2. Hébert, A., T. Guédon, T. Glatard, and B. Jaumard, High-Resolution Road Vehicle Collision Prediction for the City of Montreal. *ArXiv*, 2019.
3. Sama, M., M. Saeidi, T. Togia, and R. Kulkarni, The Effect of Negative Sampling Strategy on the Performance of the Deep Structured Semantic Model. *Proceedings of the 1st International Conference on Natural Language Processing and Information Retrieval*, 2017.
4. *Generating Negative Samples - DeepDive*, 2017, `http://deepdive.stanford.edu/generating_negative_examples`.
5. Roland, J., P. Way, and M. Sartipi, Studying the Effects of Weather and Roadway Geometrics on Daily Accidents. *Proceedings of Cyber-Physical Systems and Internet-of-Things*, 2019.
6. Athanasios, T. and Y. George, A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis and Prevention*, 2014.
7. Abdel-Aty, M. A. and A. Radwan, Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention*, 2000.
8. Khattak, A., L. Jun, and Z. Meng, Highway Safety Manual: Enhancing the Work Zone Analysis Procedure. *Transportation Research Record: Journal of the Transportation Research Board*, 2015.
9. Weng, J. and Q. Meng, Analysis of Driver Casualty Risk for Different Work Zone Types. *Accident Analysis and Prevention*, 2011.
10. See, C. F., Thesis: Crash Analysis of Work Zone Lane Closures with Left-Hand Merge and Downstream Lane Shift. *University of Kansas*, 2008.
11. Li, Y. and Y. Bai, Development of Crash-Severity-Index Models for the Measurement of Work Zone Risk Levels. *Accident Analysis and Prevention*, 2008.
12. Akepati, S. R. and S. Dissanayake, Characteristics and Contributory Factors of Work Zone Crashes. *Proceedings of Transportation Research Board 90th Annual Meeting*, 2011.
13. Ranjan, C., Extreme Rare Event Classification using Autoencoders in Keras. *Proceedings of Towards Data Science*, 2019.
14. Wilson, D., Using Machine Learning to Predict Car Accident Risk. *Proceedings of Medium - Geospatial Artificial Intelligence*, 2018.