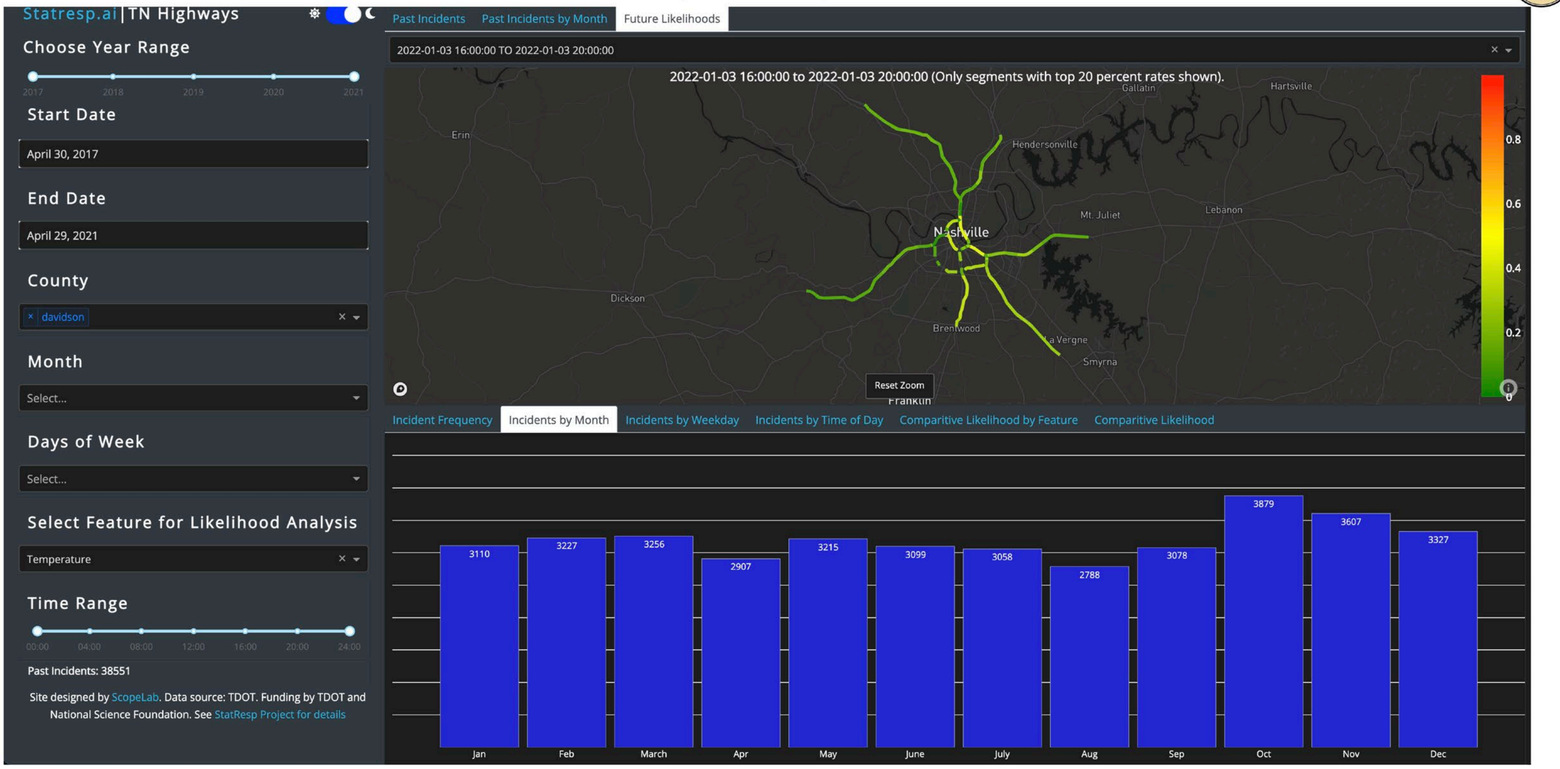
The incident analysis and prediction dashboard


Statresp.ai | TN Highways

Choose Year Range

Start Date

April 30, 2017

End Date

April 29, 2021

County

davidson

Month

Select...

Days of Week

Select...

Select Feature for Likelihood Analysis

Temperature

Time Range

Past Incidents: 38551

Site designed by ScopeLab. Data source: TDOT. Funding by TDOT and National Science Foundation. See StatResp Project for details

2022-01-03 16:00:00 TO 2022-01-03 20:00:00

2022-01-03 16:00:00 to 2022-01-03 20:00:00 (Only segments with top 20 percent rates shown).

statresp.ai

TN TDOT Department of Transportation

U.S. Department of Transportation
Federal Highway Administration

Collaborative Research Project to Coordinate the Data from the CRASH Predictive Analytics Program Between TDOT and TDOSHS

Research Final Report from Vanderbilt University | Hiba Baroud, Abhishek Dubey, Sayyed Mohsen Vazirizade, Ayan Mukhopadhyay | August 30, 2021

Sponsored by Tennessee Department of Transportation Long Range Planning

Research Office & Federal Highway Administration

VANDERBILT UNIVERSITY

DISCLAIMER

This research was funded through the State Planning and Research (SPR) Program by the Tennessee Department of Transportation and the Federal Highway Administration under ***RES #:*** 2019-02 ***Research Project Title:*** Collaborative Research Project to Coordinate the Data from the CRASH Predictive Analytics Program Between TDOT and TDOSHS.

This document is disseminated under the sponsorship of the Tennessee Department of Transportation and the United States Department of Transportation in the interest of information exchange. The State of Tennessee and the United States Government assume no liability of its contents or use thereof.

The contents of this report reflect the views of the author(s) who are solely responsible for the facts and accuracy of the material presented. The contents do not necessarily reflect the official views of the Tennessee Department of Transportation or the United States Department of Transportation.

Technical Report Documentation Page

1. Report No.
RES2019-02

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Collaborative Research Project to Coordinate the Data from the CRASH Predictive Analytics Program Between TDOT and TDOSHS

5. Report Date
September 2021

6. Performing Organization Code

7. Author(s)
Hiba Baroud, Abhishek Dubey, Sayyed Mohsen Vazirizade

8. Performing Organization Report No.

9. Performing Organization Name and Address
ISIS, Vanderbilt University,
1025 16th Ave S,
Nashville, TN, 37212

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
RES2019-02

12. Sponsoring Agency Name and Address
Tennessee Department of Transportation
505 Deaderick Street, Suite 900
Nashville, TN 37243

13. Type of Report and Period Covered
Final Report
December 2019 - July 2021

14. Sponsoring Agency Code

15. Supplementary Notes
Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.

16. Abstract

Emergency Response Management (ERM) necessitates the use of models capable of predicting the spatial-temporal likelihood of incident occurrence. These models are used for proactive stationing in order to reduce overall response time. Traditional methods simply aggregate past incidents over space and time; such approaches fail to make useful short-term predictions when the spatial region is large and focused on fine-grained spatial entities like interstate highway networks. This is partially due to the sparsity of incidents with respect to space and time. Furthermore, accidents are influenced by several risk factors that must be considered in predictive models. Collecting, cleaning, and managing multiple streams of data from various sources is challenging for large spatial areas. In this report, we highlight how this problem is being solved in collaboration with TDOT to improve ERM in TN. Working with TDOT, we have developed a novel pipeline to forecast road accidents on the interstate networks. Our pipeline, based on a combination of synthetic resampling, clustering, and data mining techniques, can efficiently forecast the spatial-temporal dynamics of accident occurrence, even under sparse conditions. Our pipeline uses data related to roadway geometry, weather, historical accidents, and traffic to aid accident forecasting. To understand how our forecasting model can affect allocation and dispatch, we improve and employ a classical resource allocation approach. Experimental results show that if our approach for forecasting road accident likelihood is employed in proactive ERM, it can reduce response times (up to 19% for 20 available HELP trucks and up to 8% on average for multiple different numbers of available HELP trucks) and the number of unattended incidents (up to 75% and 50% for mean number and maximum number of unattended accidents during a 4-hour time window, respectively) in comparison to current approaches followed by first responders. The developed pipeline is efficacious, applicable in practice, and open source. Our feature analysis also showed combination of congestion and heavy rain can increase the rate of incidents by a factor of seven while visibility and wind speed do not play a key role in prediction of the likelihood of incidents. Finally, albeit significant improvement, we provide recommendations and techniques, which we aim to investigate and apply during the next phase of the project, to enhance the performance of the model even further. The pipeline is available on https://tn.statresp.ai/.

17. Key Words
HIGHWAY CRASH, MACHINE LEARNING, PREDICTIVE ACCURACY, EMERGENCY RESPONSE

18. Distribution Statement
No restriction. This document is available to the public from the sponsoring agency at the website http://www.tn.gov/.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages	22. Price

Acknowledgement

Collaborating with Tennessee Department of Transportation (TDOT), we have been able to present our work in some peer-reviewed journals, conferences, and prestigious venues. We acknowledge the support from the TDOT for funding the research. The list of peer reviewed journal or conference papers, reports, and presentations related to this work are as follows.

Published Articles

Reports, Proceedings and Pre-prints:

Vazirizade SM, Mukhopadhyay A, Pettet G, Said S El, Baroud H, Dubey A. Learning Incident Prediction Models Over Large Geographical Areas for Emergency Response. In: 7th IEEE International Conference on Smart Computing. IEEE; 2021.

Mukhopadhyay A, Pettet G, Vazirizade S, et al. A Review of Incident Prediction, Resource Allocation, and Dispatch Models for Emergency Management. Published online 2021. https://arxiv.org/abs/2006.04200

Mukhopadhyay A, Pettet G, Vazirizade S, et al. A Review of Incident Prediction, Resource Allocation, and Dispatch Models for Emergency Management. Accid Anal Prev., *under review.*

Pettet G, Mukhopadhyay A, Vazirizade SM, Berger M, Kochenderfer M, Dubey A. Emergency Response Management Pipelines for Smart Cities.; 2020. https://statresp.ai/files/urbancomputing.pdf.

Vazirizade SM, Mukhopadhyay A, Pettet G, Said S El, Baroud H, Dubey A. Learning Incident Prediction Models Over Large Geographical Areas for Emergency Response Systems. arXiv Prepr arXiv210608307. Published online 2021.

Presentations:

Vazirizade SM, Mukhopadhyay A, Pettet G, Said S El, Baroud H, Dubey A. Learning Incident Prediction Models Over Large Geographical Areas for Emergency Response. In: 7th IEEE International Conference on Smart Computing. IEEE; 2021.

Vazirizade SM, Mukhopadhyay A, Pettet G, Said S El, Baroud H, Dubey A. Spatial Temporal Resource Demand Model for Emergency Response Management. In: Presenter, TDOT Innovation to Implementation Fair.; 2021.

Mukhopadhyay A, Vazirizade SM. Multi Agent Systems for Emergency Response. Invited Tutorial Presentation. IEEE International Conference on Smart Computing, 2021.

Executive Summary

Principled decision making in emergency response management necessitates the use of statistical models that predict the spatial-temporal likelihood of incident occurrence. These statistical models are then used for proactive stationing which allocates first responders across the spatial area to reduce overall response time. Traditional methods that simply aggregate past incidents over space and time fail to make useful short-term predictions when the spatial region is large and focused on fine-grained spatial entities like interstate highway networks. This is partially due to the sparsity of incidents with respect to the area in consideration. Further, accidents are affected by several covariates, and collecting, cleaning, and managing multiple streams of data from various sources is challenging for large spatial areas.

By leveraging the knowledge of our team in big data and machine learning, an approach was designed that collects, combines, and aggregates various datasets related to roadway geometry, weather, historical accidents, and traffic. Then, based on a combination of synthetic resampling, clustering, and data mining techniques, the proposed framework can efficiently forecast the spatial-temporal dynamics of accident occurrence, even under sparse conditions. The proposed method shows promising improvement to the current approaches followed by first responders using various metrics including real-time simulation. The model for forecasting the spatial-temporal dynamics of accident occurrence alongside strategic allocation policies can optimize and significantly improve the safety of highways.

Key Objectives

The main objective of this study was to improve the current CRASH Predictive Analytics application for highway safety patrol vehicles deployment. In other words, the goal was to show how to predict the spatial-temporal likelihood of incident occurrence for state of Tennessee, with the total area of over 100,000 sq. km.

Towards this goal, the objectives of this research are to (i) identify the best practices for data storage, integration, and maintenance infrastructure for predictive modeling, (ii) develop state-of-the-art machine learning algorithms for predicting the risk of highway incidents, and (iii) collaborate with TDOT and THP to identify best practices for model integration with existing programs

The objectives (and some key results) for the study are highlighted below:

- Evaluate all accessible information from various resources and available infrastructure that can be used for predictive modeling, and design an efficient pipeline to collect, clean, and combine, and store them (in total combining data in order of Terabytes (TB) in about a couple of hours), which can be used in future to update the final generated dataset.
- Design a pipeline using the state-of-the-art machine learning algorithms to forecast the spatial and temporal dynamics of accident occurrence, even under sparse conditions by combining assorted machine learning techniques, resulting in significant reduction of response time (up to about 4.5 minutes per incident on average).
- Develop metrics to evaluate the performance of machine learning models in predicting accidents. Importantly, it was found that conventional metrics such as correlation and accuracy might be misleading in such a sparse condition.

- Evaluate the importance of features and their combinations in predicting accident occurrence. For example, it was observed that combination of congestion and heavy rain can increase the rate of incidents by a factor of seven while visibility and wind speed do not play a key role in prediction of the likelihood of incidents.
- Provide the future path for the proposed method to enhance performance, improve robustness, and increase the reliability of the model.

Key Findings

A spatio-temporal machine learning pipeline was designed to address the problem. The pipeline used a combination of synthetic resampling, non-spatial clustering, and learning from data can efficiently forecast the spatial and temporal dynamics of accident occurrence, even under sparse conditions. To evaluate the design machine learning model, **conventional performance metrics such as accuracy, precision, recall, F1 score, Spearman correlation and Pearson correlation were used. A simulation strategy was used to measure the response time of responders and the number of unattended accidents and used it to evaluate the accuracy of predictive models** (accidents can be unattended when all resources, HELP trucks, are busy responding to other accidents). For this purpose, an allocation strategy was developed that modifies the well-known p-median problem to evaluate the performance of the models. The key findings are mentioned below:

1. **Conventional metrics such as correlation and accuracy might be misleading in such a sparse condition**. F1-score that balances the precision and recall is a much better alternative and correlates with the response performance.
2. It was observed that the proposed forecasting pipeline resulted in **significant savings in response times**. To consider the uncertainty, more than 2000 simulations were run.
3. The simulation results showed up to **19% average improvement** in response time when 20 HELP trucks were available. When the number of HELP trucks was limited to 10, the model reduced the average travel distance by responder per accident by **up to 4.5 km** (approximately more than 4.5 minutes travel time).
4. Using the prediction pipeline and proactive Emergency Response Management (ERM), responders can be placed closer to the accident-prone zones. By doing so, the travel time of the responders can be reduced, and consequently the time they are not available due to attending accidents can also be reduced.
5. An important observation was that the allocation approach, **which adds a “balancing” term to the classical p-median problem, improves resource allocation in general**; indeed, this improvement was observed across the spectrum of forecasting models used and the number of available responders.

These findings are particularly important for practitioners and first responders — while it is important to allocate resources in areas with (relatively) high historical rates of occurrence, assigning a small number of responders to cover large areas can be detrimental to the overall goal of reducing response times. Intuitively, the proposed approach penalizes additional burden on responders. However, it was also observed that a large penalty can result in increased response times. In summary, experimental results showed if our approach for forecasting road accidents is employed in proactive ERM, it can significantly reduce response times in the field in comparison with current approaches followed by first responders.

Further, a great amount of time was invested in collecting, cleaning, combining, and aggregating different datasets. For example, for weather data, datasets collected from Dark Sky[1] and Weatherbit[2], two well-known Application Programming Interfaces (APIs) for collecting historical weather data, were compared and Weatherbit was chosen. Challenges were also faced in querying elevation data from Google Cloud Elevation API. In another example, the format of the feature regarding the time of accident suddenly changed from 24-hour format to 12-hour format without keeping the am/pm label. These problems in datasets may happen, and it requires time, knowledge, and detailed investigation. Otherwise, they can compromise the performance of the model. During the last year, a lot of resources were allocated to verify the quality of the datasets. Since this process has already been investigated and conducted by our team, sharing the final combined dataset with other organizations and research teams can pave the way for other researchers and decision-makers leading to scientific and technological synergy to tackle this problem. Furthermore, collaboration with other organizations and firms can be very beneficial. For example, Google or INRIX can provide high resolution traffic data, which might not be available through any other sources.

Feature Importance

The proposed pipeline uses different features to predict the likelihood of accidents. These include congestion, precipitation, and time of the day, among others. Not all the features are equally important in their effect on the likelihood of incident occurrence. In other words, some features are more highly correlated with accident occurrence than others. The proposed approach, inspired by the concept of mutual information, evaluates the importance of each feature. Specifically, ratio of the total number of incidents given a specific feature to the frequency of that specific feature in our dataset was calculated. While results showed that visibility and wind speed are not crucial determinants of accident prediction, the combination of traffic congestion and heavy precipitation was shown to be highly correlated with accident occurrence. It was also noticed that low temperature had a slightly negative correlation to the rate of accidents. However, its combination with precipitation showed high correlation.

Key Recommendations

Even though the proposed approach showed promising results, the prediction can be improved by collecting more accident data. Having more data on incidents in a broader spatial area and longer time range allows for more advanced machine learning approaches to be applied such as model stacking techniques. On the other hand, if the available data is from another environment (for example historical accident data from another state such as California), transfer learning techniques can be used to reduce the detrimental influence of insufficient data. Moreover, incorporating spatial correlation by using graph theory will likely open the door for more accurate analysis. Additionally, this research was heavily focused on prediction while combining prediction with detection, by leveraging the crowd-sourced data platforms such as Waze[3], can also improve

[1] https://darksky.net/

[2] https://www.weatherbit.io/

[3] https://www.waze.com/

the results. The aforementioned items will be investigated in the next phase of the research. Lastly, the importance of unprecedented but powerful external factors should not be neglected. For example, the inception of COVID-19 pandemic in mid-2020 drastically changed traffic patterns. Consequently, the accident rates and patterns have unusually changed over the last year and a half.

The proposed method pushes the envelope for learning incident prediction models over large geographical areas. The results show the superiority of the performance of the proposed model compared to the current approaches followed by first responders. However, an efficient and accurate accident prediction model is one of the modules required for proactive ERM. The best prediction model without strategic allocation and optimized dispatching policies is fruitless. Therefore, other modules in EMR should be studied and improved in parallel so that the goal of TDOT, maximizing the safety of the highways, can be realized.

Table of Contents

DISCLAIMER i
Technical Report Documentation Page ii
Acknowledgement iii
    Published Articles iii
Executive Summary iv
    Key Objectives iv
    Key Findings v
    Feature Importance vi
    Key Recommendations vi
List of Tables x
List of Figures xi
Chapter 1 Introduction 1
Chapter 2 Literature Review 3
    2.1 Regression Models 3
    2.2 Random-Parameter Models 4
    2.3 Bayesian Approaches 6
    2.4 Data Mining Approaches 6
Chapter 3 Methodology 8
    3.1 Problem Formulation 8
    3.2 Challenges 8
    3.3 Data 10
    3.4 Features 12
    3.5 Approach 13
    3.6 Allocation 16
Chapter 4 Results and Discussion 19
    4.1 Model Hyper-parameters 19
    4.2 Forecasting 20
    4.3 Feature Analysis 21
    4.4 Allocation and Dispatch 23
    4.5 Implementation 32
    4.6 Discussion 35
Chapter 5 Conclusion 37

References........38
Appendices........45
1) Docker........45
2) Data Cleaning........46
3) Feature Analysis........49

List of Tables

Table I 11
Table II 28
Table III 29
Table IV 30
Table V 31
Table VI 34

List of Figures

Figure 3.1 8
Figure 3.2 10
Figure 3.3 12
Figure 3.4 13
Figure 3.5 14
Figure 3.6 17
Figure 4.1 19
Figure 4.2 22
Figure 4.3 23
Figure 4.4 24
Figure 4.5 25
Figure 4.6 26
Figure 4.7 27
Figure 4.8 33
Figure 4.9 34
Figure 4.10 35
Figure A.1 46
Figure A.2 46
Figure A.3 47
Figure A.4 48
Figure A.5 48
Figure A.6 49
Figure A.7 50
Figure A.8 51
Figure A.9 52
Figure A.10 53
Figure A.11 54
Figure A.12 55
Figure A.13 56
Figure A.14 57
Figure A.15 58
Figure A.16 59
Figure A.17 60

Chapter 1 Introduction

A constant threat that plagues humans across the globe are incidents like traffic accidents, fires, and crimes. Such incidents result in loss of life, injuries, and damage to properties and are collectively labeled as *emergencies*, which are defined as incidents that threaten public safety, health, and welfare. Consider road accidents and calls for emergency medical services (EMS) as examples. Road accidents alone account for 1.25 million deaths globally and about 240 million EMS calls are made in the U.S. each year [1]. Many such incidents make it imperative that principled methods be designed to ensure fast and effective response to incidents. At the same time, it is crucial to design infrastructure that mitigates and prevents the occurrence of such incidents. Indeed, it is well-documented that one of the most important responsibilities of federal, state, and local governments is mitigating and dealing with such events [2]. As a result, governments strive to make systematic plans, allocate resources, and take preventive measures to alleviate threats that such incidents pose.

> "An estimated 240 million calls are made to 9-1-1 in the U.S. each year."
>
> Source:
> https://www.nena.org/page/911Statistics

Emergency response management (ERM) is defined as the set of procedures and tools that first responders use to deal with incidents such as road accidents. It includes specific mechanisms to forecast incidents, detect incidents, allocate resources like ambulances, dispatch resources, and finally mitigate the post-effects of incidents [1]. Arguably, the most important component of the pipeline is to understand the spatial and temporal dynamics of incident occurrence. Gaining such an understanding can aid resource allocation and dispatch, improve the understanding of factors that cause accidents, and improve the design of safety codes. While there are several ways to analyze spatio-temporal incidents, learning data-driven forecasting models are particularly important since the fundamental goal of understanding the dynamics of accident occurrence is to aid response and dispatch. As a result, *generative* models conditional on relevant covariates are particularly relevant to the overall ERM pipeline. For example, consider a forecasting model for accident occurrence as a function of roadway speed limits. Understanding how the speed limit affects accidents helps in accurately capturing first-order effects that impact accidents, and forecasting future incidents helps shape better policy decisions pertaining to resource allocation (ambulances, for example) and response.

This report discusses a framework for predicting extremely sparse spatial temporal incidents. The project has focused on developing principled approaches to address emergency response for Tennessee, a state in the United States with a population of approximately 6.9 million and a total area of over 100,000 sq. km. While emergency response was extensively tackled by us in past collaborations with several government bodies restricted to cities [3]–[7], planning emergency response in extremely large geographical areas (an entire state, for example) is significantly more challenging. The problems are exacerbated when we limit the area of interest only to interstate highways across the state, which reduces the number of samples of positive incidents available across the road network, leading to extreme sparsity. This imbalance is particularly evident while creating forecasting models in high temporal resolution. However, the collaboration with first responders revealed that such forecasting models can be extremely beneficial for resource allocation and dispatch.

The contributions made in this research can be summarized as: 1) An efficient pipeline was designed to collect, clean, and combine, and store data. This part is the fundamental step in our pipeline since quality data with enough number of samples, spread over a long-time range, and across a large spatial region is the necessity of our machine learning engine. 2) A pipeline that can effectively forecast incidents that are sparsely scattered in space and time was developed, which can be used by ERM pipelines to reduce the average response time to accidents. It was shown that a combination of synthetic resampling and non-spatial clustering can result in the creation of accurate spatial temporal models for short-term forecasting of road accidents. 3) Unlike most forecasting models in literature, the forecasting pipeline was evaluated by measuring its effect on response times to accidents. To this end, a real-time simulation module was developed along with a novel allocation method. Through extensive simulations, it was shown how our forecasting pipeline and allocation algorithm can provide significant reduction in response times.

Chapter 2 Literature Review

A variety of approaches have been used to understand the spatial and temporal dynamics of road accidents. One of the earliest methods, known as `crash frequency analysis', uses the frequency of incidents in a specific discretized spatial area as a measure of the inherent risk the area possesses [8]. This approach also forms the basis of *hotspot* analysis [8], which is widely used in practice as a tool to visualize historical accidents and make predictions. Statistical models have also been explored in this context. The most widely used approach, Poisson regression, models the expected value of the count of incidents in each time period as a linear combination of the features. While it does not perform well on data with dispersion (where mean is not equal to the variance of the data) and sparse data, hierarchical Poisson models [9]–[13] and zero-inflated models can be used instead [14]–[17]. In recent years, data mining models such as neural networks [18]–[23] and support vector machines [24]–[26] have also been explored. The following subsections explain some of the methods more in detail.

2.1 Regression Models

One of the earliest regression models used to model incident occurrence involved multiple linear regression models with Gaussian errors [27]–[29]. However, modeling accident count by linear regression can be inaccurate, as the response variable is discrete and strictly positive. In addition, it has also been shown that linear regression models fail to model the sporadic nature of emergency incidents [30], [31]. Linear regression models with multiplicative effects have also been investigated but have shown to be inaccurate compared to other models [30]. The development of more advanced models has, therefore, made linear models almost obsolete, although occasionally it is still used due their simplicity [32]. The reason for its use is unclear. In fact, it is recommended that decision-makers carefully evaluate the shortcomings of such models before deploying them in the field. The inaccuracies of linear regression methods in the context of accident prediction is investigated in [30], [33]. While such an approach has shown performance on par with other regression models (Poisson regression, for example), it needs further validation before it is widely adopted.

The inaccuracies of linear regression and the suitability of Poisson models for count data led to the widespread use of Poisson regression for modeling incident occurrence [29]. Each incident is considered a result of an independent Bernoulli trial. Given that all the trials are generated by the same stochastic process, the series of trials can be modeled by a binomial distribution. As the number of trials becomes large and the probability of success is very small, the probability distribution over the count of incidents takes the form of a Poisson distribution [17]. To accommodate the feature vector, Poisson regression assumes that the logarithm of the expected value of the distribution is a linear combination of features. This methodology has been used extensively for emergency incident analysis [30], [31], [34]–[36].

An issue with using Poisson regression is that the expected value of the response variable (count of incidents) equals its variance. This is typically not the case with crash data, which is over-dispersed, meaning that the variance of the data is greater than its mean [17].

There are examples of incident data being under-dispersed as well [37]. Therefore, the broader argument against the use of Poisson regression is that it might not be able to model real-world crash data, which can be under-dispersed or over-dispersed. An approach to accommodate over-dispersion is to use Poisson-hierarchical models [9]. Poisson-hierarchical models (as well as Poisson models) fall under the broader category of generalized-linear models (GLM), which is a family of distributions used widely in statistics and machine learning. From this family, the Poisson-gamma (also called negative binomial) and Poisson-lognormal models are particularly relevant. The Poisson-gamma model is a Poisson distribution whose mean parameter follows a gamma distribution. It has been shown that the Poisson-gamma model fits crash data better than Poisson models, and it has been extensively used for crash prediction [10]–[12][13], [38], [39]. While the Poisson-gamma model solves the problem of over-dispersion, it performs poorly on under-dispersed data and is particularly problematic to use with small sample sizes and with data with low sample mean [40], [41]. The Poisson-lognormal model is conceptually the same as Poisson-gamma model, but it uses the lognormal distribution for the mean parameter rather than the gamma distribution [42]–[45]. The lognormal distribution is a heavy tail distribution and provides more flexibility for over-dispersion. Recently, the Poisson-inverse-gamma model has been used in crash modeling [46]. However, such models do not have closed-form maximum likelihood estimation (MLE) solutions unlike the Poisson-gamma models [47].

Despite the success of Poisson and Poisson-hierarchical models, a common shortcoming is that both models fail to adequately handle the prevalence of zero counts in crash data [17]. A remedy to this problem is to use zero-inflated models, and both zero-inflated Poisson and zero-inflated Poisson-gamma models have been used to model accident data [14]–[16]. Zero-inflated models can be described as having dual states, one of which is the *normal* state, and the other the *zero* state. The excess zeros that cannot be explained by standard count-based models can then be considered to have arisen due to the presence of a separate state. Zero-inflated models result in improved statistical fit to accident data. However, it was noted [17] most prior works justify the use of zero-inflated models by improved likelihood, and therefore automatically assume that crash data is generated by a dual-state process (except [30], which uses a zero-inflated model to justify misreporting of incidents). Through empirical data and simulations, they show that excess zeros could arise due to various other factors like low traffic exposure and the choice of spatial and temporal scales by the model designer. As a result, it is not clear if the statistical backing to using dual-state models is accurate or not.

2.2 Random-Parameter Models

Accounting for unobserved heterogeneity (i.e., factors affecting incident frequency but not captured in the data) has dominated recent statistical modeling development, with random-parameter (RP) models being among the most widely used approaches [48]. Unobserved heterogeneity introduces a variation in the effect of observed variables on the outcome. The outcome is typically the likelihood and severity of a crash. For example, a highway's design speed limit is a commonly used variable in the prediction of the likelihood of crashes. However, this may introduce unobserved heterogeneity if the vehicle's actual speed is not considered which may be different than the design speed limit across different drivers. Environmental conditions are also commonly used to explain crash occurrence and severity such as time of the day and weather variables. However, the same amount of precipitation may lead to different outcomes in the likelihood and severity of accidents depending on the geographic area and the different ways drivers respond to adverse conditions.

Additionally, unobserved heterogeneity can result from the spatial or temporal aggregation of accidents. Since these events are rare, they are often aggregated over time (e.g., number of accidents per 4 hours) or space (e.g., number of accidents per road segment) before they are modeled. The lack of consideration for unobserved heterogeneity will lead to biased estimates because the effect of an observed variable will be the same across all observations for a particular instance [48]. RP models address heterogeneity by allowing the estimated parameters to vary across observation according to a continuous distribution. A significant portion of RP models in the literature are based on the assumption that random parameters follow a distribution with a common mean and no mutual dependence [49], [50]. However, lack of consideration of cross-correlation and mutual dependence can lead to biases in the estimation of parameter variances [51]. A few recent studies have considered cross-correlated RP models and compared their performance to fixed-parameters and uncorrelated RP models. The correlated RP negative binomial model resulted in an improved log-likelihood compared to the fixed-parameters model [52] and better statistical performance and predictive power compared to the uncorrelated model [53]. In another study, correlated RP Tobit model was shown to outperform both fixed-parameters and uncorrelated RP Tobit models [54]. However, these results are still not conclusive as other studies have found the relative statistical performance between uncorrelated and correlated RP count models to be comparable [55]. Therefore, additional research is needed to determine the advantages of correlated RP models.

In addition to cross-correlations and improved statistical performance, another advantage of using correlated RP models is the ability to account for the heterogeneous effects of covariates across roadway segments as they apply to crash frequency analysis on multilane highways [55]. While the focus of this section is on RP models as they are the most adopted methods, it is worth noting that other approaches have been developed to address unobserved heterogeneity (see the work by [48] for an extensive review). For instance, latent-class (finite mixture) models seek to identify groups of observations having homogeneous variable effects [56]. These models do not require a parametric assumption for the distribution of estimated parameters like RP models; however, they still impose a parametric model structure and can be computationally intensive. To account for the variation at both the group and individual observation levels, RP models within each class have been used with mixture models [57]. Other approaches address specific heterogeneity issues such as Markov-switching models which have been used for time-dependent unobserved heterogeneity [58]. Such a form of heterogeneity can be caused by time-varying factors such as traffic and weather conditions or when the accidents are aggregated over a certain period.

2.3 Bayesian Approaches

Bayesian methods [59], [60] are often used for parameter estimation. Such models result in a distribution over parameters rather than point estimates, which can result in greater robustness to outliers and small sample sizes [61]. The empirical Bayes method (also known as maximum marginal likelihood) has been used in traffic engineering [62]–[65]. Bayesian modelling techniques have also been used to assess potential risk factors of spatial regions [66], [67] and to estimate expected crash frequencies [68].

Hierarchical Bayesian estimation of safety performance models have also been explored over the last two decades [40], [43], [44], [69]–[71]. Recently, the Poisson-gamma and Poisson-lognormal models have also been estimated using Bayesian [10]–[13], [38], [39], [42], [45], [46], [72] methods. A caveat regarding Bayesian models is that the crucial choice of priors in the predictive models. The underlying information for designing priors might be available from previous models, engineering judgement, etc., and prior distributions can also be chosen to be non-informative or weakly informative. An important investigation in this context, specifically regarding crash prediction, has been done [73], who study the performance of various Bayesian multivariate spatial models with different prior distributions. It has also been shown that using non-informative priors may result in a high bias for the dispersion parameter in models, especially with small sample sizes [74].

2.4 Data Mining Approaches

With improved sensor technology and easier storage, data-mining methods have successfully been used for crash prediction. This has also resulted in the creation of richer feature sets, which aid the performance of such methods. These days, various ways of collecting data are employed to gather data about traffic, traffic incidents, and the features related to that. By doing so, big data sets are available, which is the requirement of data mining methods. Random forests [75], [76], support vector machines [24]–[26] and neural networks [18]–[21] have recently been used to model crashes. Bayesian neural networks have also been explored, which address over-fitting of neural-networks in crash modeling [77]. Deep learning techniques have also been used in various studies [22], [23]. One of the models that may be of interest to practitioners was developed using a spatio-temporal convolution long short-term memory network (LSTM) to predict short-term crash risks, including propagation of traffic congestion [78]. While the network structure was a combination of various complex networks, the accuracy of hourly predictions was limited, which highlights the inherent difficulty of predicting crash frequency at low temporal and spatial resolutions. It also makes a case against the use of complex models in this domain because they are harder to generalize.

Ensemble methods use multiple trained models to improve prediction compared to what can be obtained from any of the individual models. While the simplest approach is averaging the prediction of assorted models, a better approach is to intelligently aggregate the prediction of these models by employing another learning algorithm, which is called model stacking. Big data and surge in availability of computational resources have paved for more sophisticated approaches such as model stacking to be used in incident prediction. Various stacking models with different numbers of layers and assorted types of models [79]–[85] have been used during recent years to predict and detect incidents. The main caveats of using ensemble models are overfitting and the size of the data required for testing and training.

Chapter 3 Methodology

The problem setting (the Interstate Highway network of the state of Tennessee) used in this project is shown in Figure 3.1. The focus on some of the highways and not all, as driven by the availability of the historical incident data. In the next phase, the prediction would be extended to all highways in the state of Tennessee.

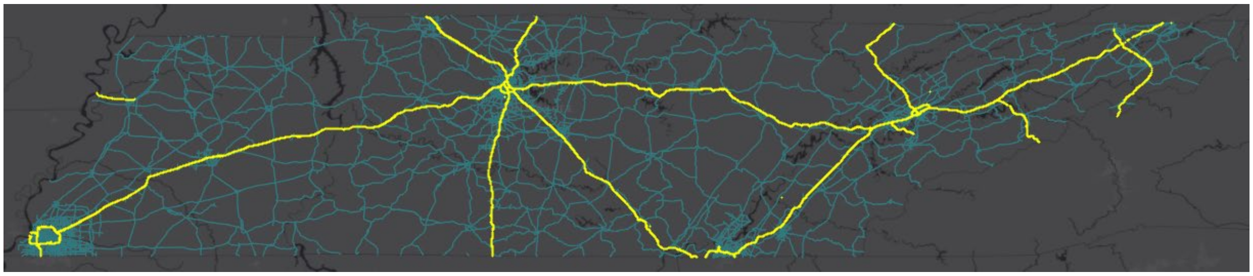


Figure 3.1 Blue (and thin) lines represent TN's roadway network. Yellow (and thick) lines represent interstate highway segments under the jurisdiction of TDOT and are the area of study for this study

The following subsections describe the approach and the mathematical formulation of the problem.

3.1 Problem Formulation

Consider a spatial area of interest S, in which incidents (like accidents) occur in space and time. The decision-maker observes a set of samples (possibly noisy) drawn from an incident arrival distribution. These samples are denoted by $(s_1, t_1, k_1, \omega_1), (s_2, t_2, k_2, \omega_2), \ldots, (s_n, t_n, k_n, \omega_n)$ where s_i, t_i and k_i denote the location, time of occurrence, and reported severity of the i^{th} incident, respectively, and $\omega_i \in \mathbb{R}^m$ represents a vector of features associated with the environment defined by the location and time of the incident. We refer to this tuple of vectors as D which denotes the input data that the decision-maker has access to. The vector ω can contain spatial, temporal, or spatio-temporal features and it captures covariates that potentially affect incident occurrence. For example, ω typically includes features such as weather, traffic volume, and time of day (we describe all features in detail later in the text). The most general form of incident prediction can then be stated as learning the parameters θ of a function over a random variable X conditioned on ω. We denote this function by $f(X \mid \omega, \theta)$. The random variable X represents a measure of incident occurrence such as a *count* or *presence* of incidents during a specific time-period. The goal of the incident prediction problem is to find the *optimal* parameters θ^* that best describe D. This can be formulated as a MLE problem or an equivalent empirical risk minimization (ERM) problem.

3.2 Challenges

The problem described in previous section is hard due to the following challenges.

Irregular incident occurrence: It is well-established in literature that predicting road accidents is extremely difficult due to inherent randomness of accidents and spatially varying factors [1], [86]. While accidents are affected by various features, it is difficult to take all determinants into account while designing forecasting models. For example, consider the condition of a specific road. It is difficult to observe such features in real-time, thereby resulting in unobserved heterogeneity in the likelihood of incident occurrence across a large spatial region. Indeed, sophisticated models have underperformed in predicting accidents. For example, an approach particularly important to practitioners was developed by Bao et al. [23], who used a spatio-temporal convolution long short-term memory network (LSTM) to predict short-term crash risks. While the network structure was a combination of various complex sub-networks, the accuracy of hourly predictions was limited, highlighting the inherent difficulty of predicting crash frequency at low temporal and spatial resolutions.

Sparsity: It is also crucial to consider the frequency of incident occurrence. While the frequency of road accidents is alarming, incidents are extremely sporadic when viewed from the perspective of total time and space in consideration. For example, there were a total of approximately 78,000 road accidents reported between 2017-2020 on interstate highways in Tennessee. Now, consider the goal of learning the dynamics of incident occurrence. While historical data can be studied using *hotspots* to improve policy, short-term forecasting models are important for deploying ambulances, HELP trucks, and other emergency responders. Based on conversations with first responders, it was found that short-term deployment often occurs several times in a day, the most common frequency being once every four hours. Considering a total of about 5,000 road segments and time slots of four hours, the data shows >99% sparsity. This challenge is represented schematically in Figure 3.2 by randomly selecting 180 road segments for April 2019 and 180 four-hour time slots. Each pixel in the matrix denotes the presence (white) or absence (black) of an accident in a segment (denoted by rows) in a span of four hours (denoted by columns). Most of the matrix consists of black pixels 99.8%, making such problems extremely difficult from the perspective of data-driven modeling. In comparison, previously studied statistical models can be used to predict incidents in small urban areas [3]–[5] (such situations typically exhibit <90% sparsity).

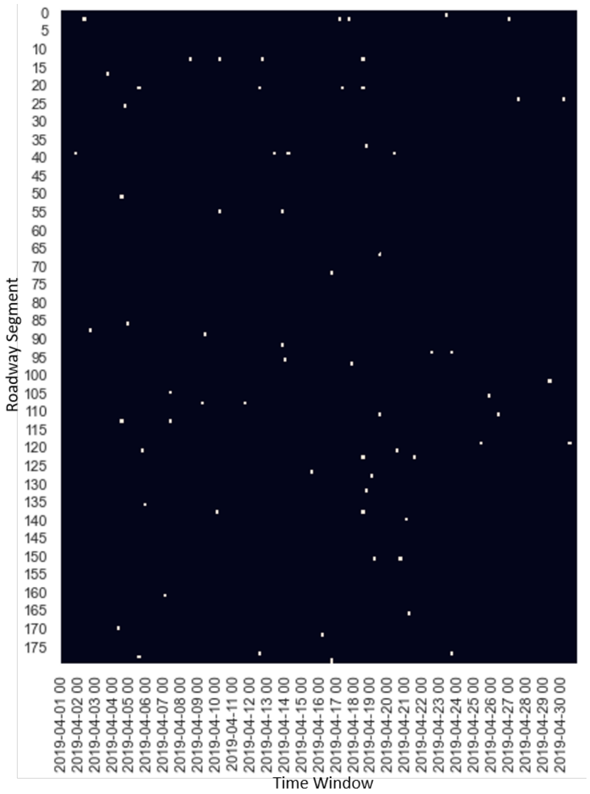


Figure 3.2 Schematic overview of the sparsity of accident occurrence across space and time. The figure shows randomly selected 180 road segments for four-hour time windows in April 2019

Data Integration: Road accidents are affected by many determinants which can be spatial, temporal, or spatial temporal in nature. For example, the geometry of a specific road segment does not change over time and is an example of a spatial feature. Time of day, on the other hand, is an example of a temporal feature. Some features can be affected by both space and time; for example, traffic congestion in a specific area is determined by the spatial location of the area as well as time of day. For predicting accidents in large geographic areas, it is challenging to collect, clean, understand, and analyze data from different sources and integrate them into models for incident prediction.

3.3 Data

The covariates used and their sources are described in Table I. It is crucial to reiterate the importance of this stage in real-world machine learning pipelines; in fact, the availability of multiple streams of data has been noted as being particularly important for predicting accidents [1].

Table I
Data Features, Size and Sources

Dataset	Range	Size	Rows	Feature	Source	Freq.	Type	Definition
-	-	-	-	Time of day	derived	-	Temporal	We divide each day into six 4-hour time windows.
-	-	-	-	Weekend	derived	-	Temporal	A binary feature that denotes weekdays.
Incident	02/01/2017 to 05/01/2020	21MB	80K	Past Incidents in the last window	derived	-	Spatio-temporal	Number of incidents on the segment in the last time window of 4 hours
				Past Incidents in a day	derived	-	Spatio-temporal	Number of incidents on the segment in the last day
				Past Incidents in a week	derived	-	Spatio-temporal	Number of incidents on the segment in the last week
				Past Incidents in a month	derived	-	Spatio-temporal	Number of incidents on the segment in the last month
Weather	02/01/2017 to 06/01/2020	300MB	1.4M	Visibility	Weatherbit	1 hour	Spatio-temporal	A measure of the distance at which an object or light can be clearly discerned.
				Wind Speed	Weatherbit	1 hour	Spatio-temporal	Speed of wind
				Precipitation	Weatherbit	1 hour	Spatio-temporal	Amount of precipitation.
				Temperature	Weatherbit	1 hour	Spatio-temporal	It is the reported temperature.
Traffic	04/01/2017 to 12/01/2020	1.2TB	30B	Congestion	derived	5 minutes	Spatio-temporal	Congestion is the ratio of the difference between free flow speed and the current speed to free flow speed
				Free Flow Speed	INRIX	5 minutes	spatial	The speed at which drivers feel comfortable if there is no traffic and adverse weather condition.
				Traffic Confidence	INRIX	5 minutes	Spatio-temporal	A confidence score regarding the accuracy of the traffic data (we collect this directly from INRIX).
Roadways	Static	81MB	80K	Lanes	INRIX	static	Spatial	Number of lanes for a roadway segment.
				Miles	derived	static	Spatial	Length of a roadway segment.
				iSF	derived	static	Spatial	Inverse scale factor which represents the the curvature of a roadway segment.

In the appendices section, some of the main challenges in the data collection are presented along with the approaches that were taken to address them.

3.4 Features

This section describes the features we extract from the base data Table I and use as covariates in our pipeline.

Roadway Information: To learn a predictive model for accidents over a graph of roadways, it is imperative to first define the edges and the vertices of the graph. Roadway information was collected from INRIX [87], a private entity that provides location-based data and analytics, such as traffic and parking, to automakers, cities, and road authorities worldwide. Information was retrieved for about 80,000 roadway segments in the state of Tennessee, out of which about 5,000 are interstate highway segments. Data about static features associated with each segment were also retrieved that are immutable (relatively) over time. For example, for each road segment, information was collected about the number of lanes, length, and coordinates. To evaluate how roadway shape affects accidents, a feature called the *inverse stretch factor* (iSF) was introduced, that represents the curvature of road segments. An example for calculating iSF is shown in Figure 3.3. For the segment in consideration (between points A and B), iSF can be calculated as the length the straight line $\overline{AB}$ divided by that of the curve $\widehat{AB}$.

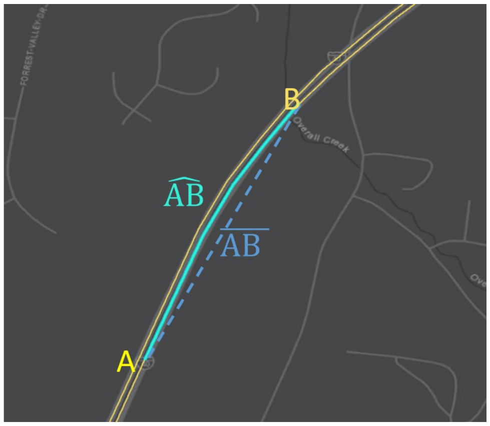


Figure 3.3 A combination of the length of a curve (AB)ˆand the shortest path between the two ends of the curve ¯AB can be used to denote its curvature

Traffic: The correlation of traffic and road accidents is well-established [1]. We collected traffic data for each of the road segments through INRIX at a temporal resolution of 5-minute intervals for about three years. Specifically, information was retrieved regarding the free flow speed of traffic, the estimated current speed of the vehicles, and the confidence scores of the estimates. Effective congestion can be calculated from data as the ratio of the difference between the free flow speed and the current speed to the free flow speed.

Weather: Weather is inherently spatial temporal, and can play an important role in accident rates [1]. We collected hourly weather data (temperature, precipitation, visibility, and wind) from 40 different weather stations in and around the state of Tennessee. The locations of the stations are shown in Figure 3.4. To use weather data to forecast accidents on a given road segment, the weather station that is the closest to that particular segment was used.

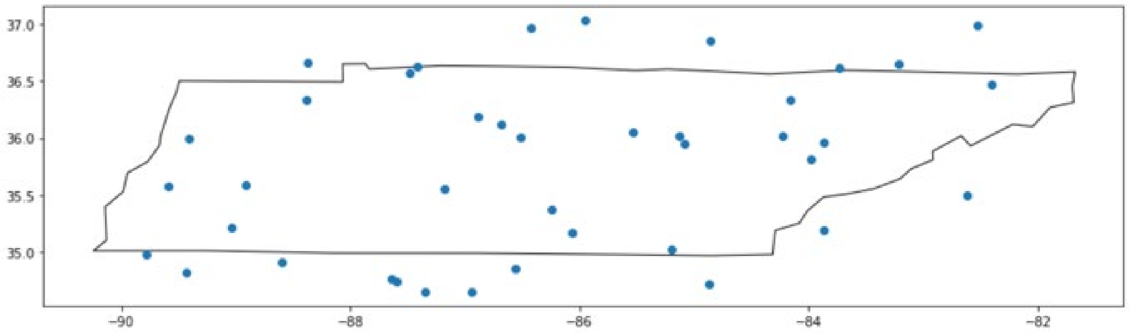


Figure 3.4 Location of the weather stations

Incidents: Every accident reported in Tennessee from January 2017 to May 2020 was considered. Incident data for this project is provided by the TDOT and consists of approximately 78,000 accidents. The accuracy of the incident data was verified with the Enhanced Tennessee Roadway Information Management System (E-TRIMS).

In summary, the following features were used: time of day, weekend, past incidents in the last window, past incidents in a day, past incidents in a week, past incidents in a month, visibility, wind speed, precipitation, temperature, congestion, free flow speed, traffic confidence, lanes, miles, and inverse stretch factor (iSF).

3.5 Approach

The approach to predict roadway accidents in space and time is shown in Figure 3.5. To begin with, road segments that exhibit no accidents or extremely small number of accidents over the temporal period in consideration (about three years) were filtered out. The analysis is done on 77% of the observed accidents with a total sparsity of 98%. Recall the fundamental goal of the project is to learn a function f that outputs the likelihood of incident occurrence on a road segment conditional on a set of features. A straightforward way to do so is to learn a separate model over each segment. However, such an approach results in overfitting; each segment contributes a relatively small amount of data which ignores structural similarities between patterns of incident occurrence across the entire spatial region in consideration. The other approach is to learn one model for the entire area. However, a universal model fails to capture any heterogeneity that is not explicitly modeled in the feature space. To balance these considerations, segments that observe similar patterns for incident occurrence were identified. While it is possible to identify distinct spatial regions (hotspots) and learn a separate model for each area, it is possible that there exists generalizable information in the entire area that is spatially invariant. To do so, common areas were identified irrespective of spatial contiguity by clustering all the available segments based on their frequency of incident occurrence. In this study, the well-known k-means algorithm [88] was used to group the segments into distinct clusters.

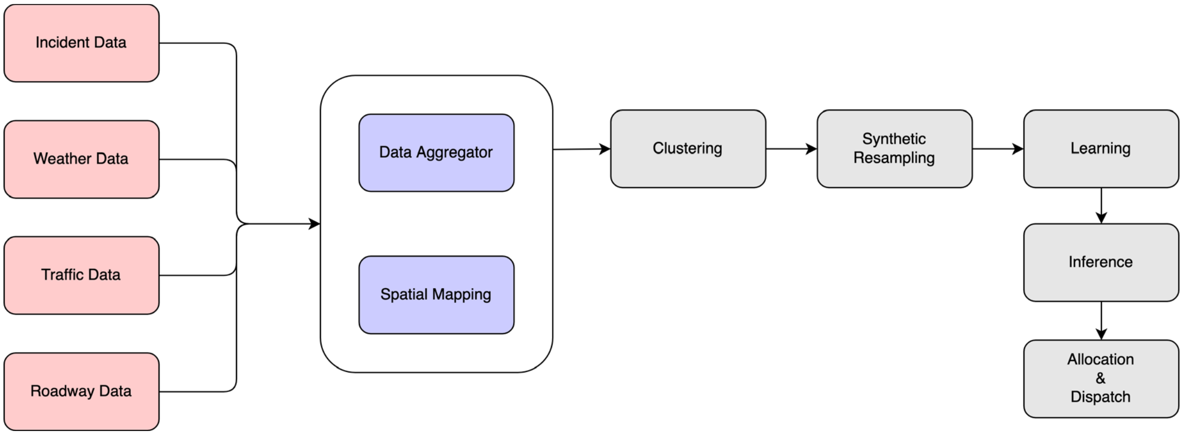


Figure 3.5 Overview of the proposed approach. Spatial temporal information extracted from a variety of data sources were combined and segments were clustered to focus on heterogeneity not explicitly modeled in the feature space. Then, synthetic sampling was used to address sparsity and finally, statistical and algorithmic models were learned on on incident occurrence

Given clusters of roadway segments that share similar patterns of spatial-temporal incident occurrence, learning the patterns is still challenging due to the sparsity of the data. To address this concern, synthetic under-sampling and over-sampling were performed to balance the data. However, naive synthetic sampling performs poorly in this problem setting since the relative frequencies of incident occurrence are markedly different among the clusters. Therefore, it is impractical to 'balance' data in each cluster in the same manner. To alleviate this, the proposed approach starts with the cluster with the highest frequency of incident occurrence (cluster A, say) and performs synthetic sampling such that the number of positive data points (spatial segments in temporal windows that have accidents is the same as the number of negative data points (spatial segments in temporal windows that do not have accidents). Then, synthetic sampling is performed in the other clusters such that the ratio of accidents occurring for any given cluster (cluster B, say) to the frequency in A is the same as in the original dataset[4]. Clustering and synthetic sampling provide the foundation for learning spatial temporal forecasting models over accident occurrence. The following well-known models were used to this end.

Logistic Regression (*LR*): There are two classes of approaches that can be used to forecast the chances of accidents on road segments. First, one can try to model the count of accidents as a binary variable and use well-known count-based regression models like Poisson regression, zero-inflated Poisson regression, and negative binomial regression. The other approach is to treat the occurrence of accidents as a dichotomous output and model the likelihood that any accidents occur. Logistic regression was used for this problem setting, which models the log-odds of the probability of incident occurrence as a linear combination of the features ω.

[4] we also show results without synthetic sampling and clustering.

Zero-Inflated Poisson (ZIP): Count-based models were also used to model accident occurrence conditional on spatial temporal features. While Poisson regression has been widely used to model accident occurrence, hierarchical Poisson models and zero-inflated models have demonstrated significantly improved predictive power [15]. ZIP models can be described as having dual states, one of which is the normal state, and the other the zero state [89].

Random Forests (RF): RF classifiers are a decision tree ensemble method where each tree is constructed from independently bootstrapped samples [90]. They reduce model variance and are less likely to overfit compared to standard decision trees due to bootstrap aggregation and the use of a random selection of features to split nodes when constructing each tree (called 'feature bagging'). In addition to synthetic sampling, random forests can address sparsity using the Balanced Random Forest method [91]. This works by assigning weights to each class inverse-proportionally to their frequency in the dataset, giving a heavier penalty to misclassifying the minority class.

Neural Networks (NN): Finally, simple artificial neural networks were also used to learn a model over incident occurrence. Neural networks consist of a set of layers, each of which further consists of neurons or computing units. The output of each layer is fed as input to the next layer [92]. Each neuron uses a non-linear function (called the activation function) of the sum of its inputs and produces an output. The network can be trained by stochastic gradient descent. We use fully connected layers in this study. An important note to practitioners is the non-interpretability of neural networks can be a barrier when deploying systems in the real-world that affect government policies.

It is natural to compare forecasting approaches through metrics like likelihood values on test data, error rates, precision, and recall. However, conversations with first responders revealed that it is particularly beneficial for them to understand if forecasting models can rank roadway segments based on risk. This is intuitive since accurately forecasting the risk at each segment relative to other segments is important for allocating resources. Therefore, besides standard statistical metrics (accuracy, precision, recall, F1-score), the correlation of each model's marginal accident likelihood distribution over space with the real accident distribution is also reported. Specifically, both Pearson and Spearman correlation values are presented in this report.

3.6 Allocation

The primary purpose of incident prediction models is to make informed resource allocation decisions. However, prior literature rarely evaluates ERM pipelines in their entirety. As a result, the project aimed to provide a simple and flexible allocation approach based on our incident prediction model to evaluate that. This approach is employed to distribute the responders. Therefore, the impact that incident models have on response time outcomes was evaluated. The evaluation process was guided by the following steps:

1) Understanding existing policies: Through our collaboration with first responders, it was first understood how emergency resources are allocated and deployed in practice. This approach is described below and used as our baseline.
2) Resource Allocation: In practice, discrete location models like the well-known p-median formulation [93]–[96] are widely used to allocated emergency resources. A shortcoming of such approaches is that the service time of resources (for example, the time that ambulances are busy responding to accidents) is not taken directly into account in the allocation process. For this project, a novel modification to the p-median problem was introduced that bridges this crucial gap.
3) Evaluation: Using the proposed allocation model, the performance of existing prediction models and our pipeline was evaluated by creating a black-box simulator that imitates emergency response.

Resource allocation is often based on identifying hotspots of incident occurrence. First, a map based on historical accidents is created. Then a group of experienced engineers determine the location of the responders; typically, responders are placed in areas with the highest historical accident rates. The allocation formulation we use is based on the p-median problem, which is commonly applied to ambulance allocation. The objective of the standard p-median problem is to locate p facilities (i.e., responders) such that the average demand-weighted distance between edges and their nearest facility is minimized. One shortcoming of the p-median formulation is that it does not account for responders becoming unavailable when attending to incidents. To address this, the standard p-median was modified by adding a balancing term to the objective function. Intuitively, this balancing term penalizes responders that cover disproportionately large demand compared to other facilities, encouraging multiple responders to congregate near high demand areas. This effect is schematically demonstrated in Figure 3.6. In the figure, the values in the cells correspond to the chance of accident occurrence for the location and the green points show the allocated locations of responders (*p*=2 in this case). By considering $\alpha = 0$ (alternative a), the problem is equivalent to the simple p-median formulation, which seeks to minimize the weighted distance between allocations and points of demand. However, by increasing α (alternative b), the optimizer seeks to avoid assigning high risk cells to a single responder. Formally, the following optimization problem is solved:

$$\min \sum_{i=1}^{|E|} \sum_{j=1}^{|L|} a_i \, d_{i,j} \, Y_{i,j} b_j$$

$$\text{s.t} \sum_{j=1}^{|L|} Y_{i,j} = 1, \forall i \in \{1, \dots, |E|\}$$

$$\sum_{j=1}^{|L|} X_j = p$$

$$Y_{i,j} \leq X_j, \forall i \in \{1, \dots, |E|\}, \forall j \in \{1, \dots, |L|\}$$

$$Y_{i,j}, X_j \ \in \{0, 1\}, \forall i \in \{1, \dots, |E|\}, \forall j \in \{1, \dots, |L|\}$$

where E is the set of demand nodes, i.e., roadway segments, L is the set of possible responder locations, p is the number of responders to be located, a_i is the likelihood of accident occurrence on edge $e_i \in E$ and $d_{i,j}$ is the distance between edge $e_i \in E$ and location $l_j \in L$. $Y_{i,j}$ and X_j are two sets of decision variables; X_j = 1 if a responder is located at $l_j \in L$. and 0 otherwise, and $Y_{i,j}$ = 1 if edge $e_i \in E$ is covered by a responder located at $l_j \in L$ (i.e., the responder at l_j is the nearest placed responder to e_i) and 0 otherwise. The balancing term we add is denoted by $b_j = \left(\frac{\sum_e a_e Y_{e,j}}{\sum_e a_e}\right)^{\alpha}$, and represents the proportion of total demand covered by a responder located at j. The influence of the balancing term is controlled by the hyper-parameter α; intuitively, as α increases, responders are more `tightly packed' around high demand areas, and if α = 0 our formulation reduces to the standard p-median formulation.

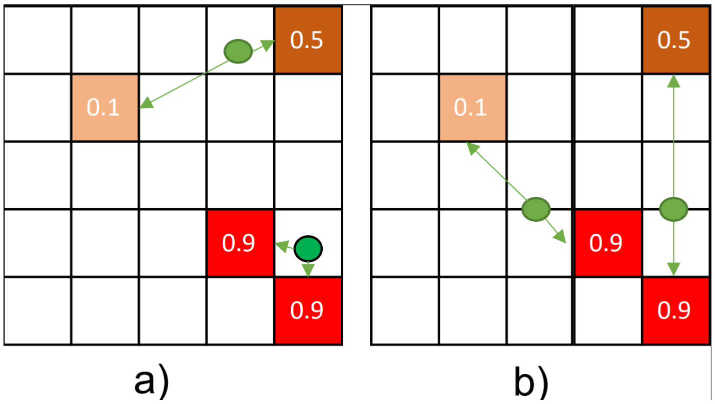
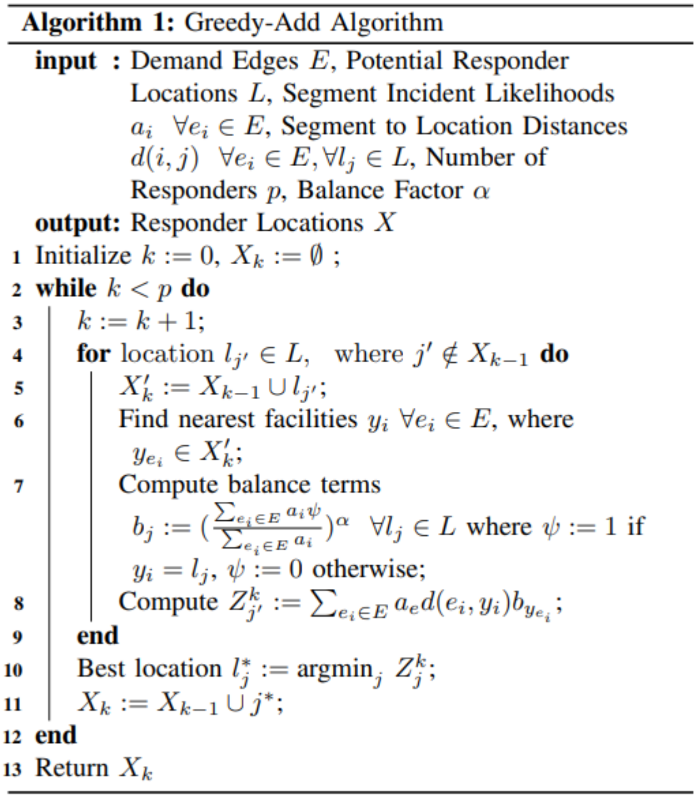


Figure 3.6 Illustrating the impact of α a) standard p-median (α =0). b) modified p-median with α >0. Notice as α increases responders (green dots) are tightly packed around high demand areas

The p-median problem is known to be NP-hard on general networks [97]. Therefore, heuristic methods are employed to find approximate solutions in practice. For this project, the Greedy-Add algorithm [98] was used to optimize the locations of responders. The algorithm is shown in Algorithm 1. First, the iteration counter k is initialized to 0 and the set of allocated responder locations X_k to the empty set (step 1). Then, as long as there are responders awaiting allocation, the following loop is iterated through: (1) update counter k current iteration (step 3) (2) for each potential location not already in the allocation, compute the modified p-median score of the allocation which includes the potential location (steps 5 - 8), and (3) find the location that minimizes the modified p-median score (step 10) and add it to the set of allocated responder locations (step 11). While myopic, this algorithm is scalable to large allocation problems.



```
Algorithm 1: Greedy-Add Algorithm
  input : Demand Edges $E$, Potential Responder
          Locations $L$, Segment Incident Likelihoods
          $a_i \ \forall e_i \in E$, Segment to Location Distances
          $d(i,j) \ \forall e_i \in E, \forall l_j \in L$, Number of
          Responders $p$, Balance Factor $\alpha$
  output: Responder Locations $X$
1  Initialize $k := 0$, $X_k := \emptyset$ ;
2  while $k < p$ do
3      $k := k + 1$;
4      for location $l_{j'} \in L$,  where $j' \notin X_{k-1}$ do
5          $X'_k := X_{k-1} \cup l_{j'}$;
6          Find nearest facilities $y_i \ \forall e_i \in E$, where
            $y_{e_i} \in X'_k$;
7          Compute balance terms
            $b_j := (\frac{\sum_{e_i \in E} a_i \psi}{\sum_{e_i \in E} a_i})^\alpha \ \ \forall l_j \in L$ where $\psi := 1$ if
            $y_i = l_j$, $\psi := 0$ otherwise;
8          Compute $Z^k_{j'} := \sum_{e_i \in E} a_e d(e_i, y_i) b_{y_{e_i}}$;
9      end
10     Best location $l^*_j := \text{argmin}_j \ Z^k_j$;
11     $X_k := X_{k-1} \cup j^*$;
12 end
13 Return $X_k$
```

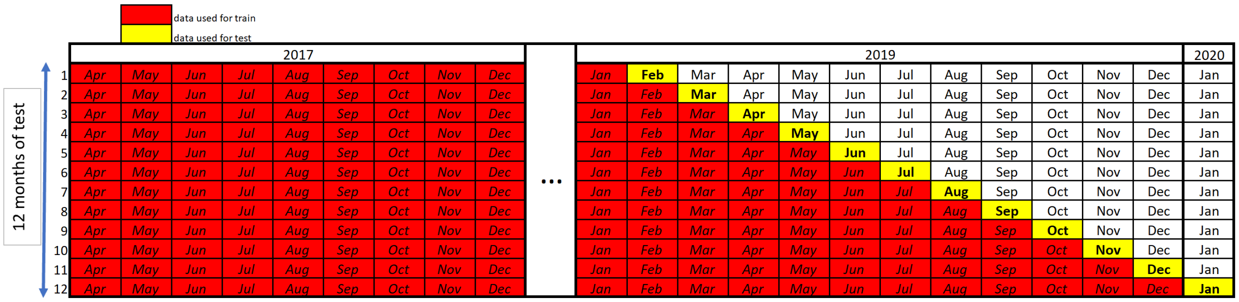
Rather than restricting responders to the roadway segments *E*, they are allowed to be located anywhere across the state in the proposed setting. To accomplish this, the set of possible responder locations *L* was considered as a grid of spatial cells over Tennessee. Each grid cell is created of the size of 0.1 degrees latitude by 0.1 degrees longitude, which is approximately 9km x 11km in Tennessee. This results in 1445 possible locations across the state. The center of each cell is used when calculating the distance between it and each edge in *E*.

Given an allocation of responders, response to real incidents was simulated to evaluate the efficacy of our model. Response to emergency incidents is typically greedy; the closest available responder to the scene of the incident is dispatched to attend to it. This is a direct consequence of the critical nature of the incidents that emergency responders address. A simulator that imitates greedy dispatch was used to evaluate the performance of different predictive models.

Chapter 4 Results and Discussion

ERM necessitates the use of models capable of predicting the spatial-temporal likelihood of incident occurrence. These models are used for proactive stationing to reduce overall response time. Traditional methods simply aggregate past incidents over space and time; such approaches fail to make useful short-term predictions when the spatial region is large and focused on fine-grained spatial entities like interstate highway networks. This is partially due to the sparsity of incidents with respect to space and time. Furthermore, accidents are affected by several covariates. Collecting, cleaning, and managing multiple streams of data from various sources is challenging for large spatial areas. In this section, results using our method for forecasting the spatio-temporal dynamics of accident occurrence are discussed based on various performance metrics as well as simulation. Recommendations and techniques, which aim to investigate and apply during the next phase of the project are also discussed. Also, the benefits to TDOT on the potential implementation are also mentioned.

To evaluate our models, we use actual historical incident data, roadway geometry, traffic data, and weather data. Each model was trained based on a rolling temporal window as shown in Figure 4.1.



🟥 data used for train

🟨 data used for test

12 months of test

	2017										2019												2020
1	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	**Feb**	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan
2	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	*Feb*	**Mar**	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan
3	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	*Feb*	*Mar*	**Apr**	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan
4	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	*Feb*	*Mar*	*Apr*	**May**	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan
5	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	*Feb*	*Mar*	*Apr*	*May*	**Jun**	Jul	Aug	Sep	Oct	Nov	Dec	Jan
6	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*	…	*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	**Jul**	Aug	Sep	Oct	Nov	Dec	Jan
7	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	**Aug**	Sep	Oct	Nov	Dec	Jan
8	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	**Sep**	Oct	Nov	Dec	Jan
9	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	**Oct**	Nov	Dec	Jan
10	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	**Nov**	Dec	Jan
11	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	**Dec**	Jan
12	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*		*Jan*	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*	*Oct*	*Nov*	*Dec*	**Jan**

Figure 4.1 Each row describes a test case. The months in yellow (also the text is bold) are the forecasting target. The months in red (also the text is italic) are used to train the models for prediction for the specific row

4.1 Model Hyper-parameters

Hyper-parameters for each model were tuned by cross-validation. For models based on random forests and neural networks, the architecture was kept fixed based on the largest training sample; classification thresholds were tuned for every training window based on a validation set. The model parameters are described below:

RF: Each random forest consisted of 250 decision trees. Gini impurity was used to measure the quality of a node split, and considered $\sqrt{|w|}$ random features for each split, where w is the total number of features. The following hyper-parameters were tuned for each model: the maximum depth of each tree, the minimum number of observations in a node required to split it, and the minimum number of samples required to be at a leaf node to split it's parent.

NN: A sequential architecture was used with fully connected layers. We used a total of three hidden layers. The size of the first layer equals twice the number of input features

w (the number of neurons in the input layer). The second and third layers consisted of neurons equal to the size of the input layer. The output layer consisted of a single neuron. We used the 'ReLU' activation function [99] for all hidden layers and the sigmoid activation function for the output layer. The cross-entropy loss between true labels and predicted labels was minimized by using the Adam algorithm [100] for training the network.

Clustering: The k-means algorithm [88] was used to group the segments into clusters (k=2 was set for this analysis). A higher value for k rendered an extremely small number of segments in some of the clusters, thereby hampering overall performance. Naturally, it is recommended that practitioners tune all hyper-parameters based on the specific dataset in consideration.

4.2 Forecasting

The performance of the forecasting pipeline is presented below. The following abbreviations are used for brevity: LR (logistic regression), NN (neural networks), RF (random forests), ZIP (zero-inflated Poisson), RUS (random under sampling), ROS (random over sampling), NoC1 (No clustering) and KM2 (k-means clustering). The proposed baseline was based on the actual forecasting model that aids first responders in Tennessee. This baseline is referred as the naive model. The naive model essentially created an empirical distribution based on historical incident data. Then, given a segment, a specific point in time, and the set of covariates induced by them, a realization of incident occurrence was sampled from the empirical distribution conditional on the covariates. Results are presented for each of the approaches in Table II. To understand the role and efficacy of each component of the pipeline, results are presented with and without synthetic resampling and clustering.

The major observations are as follows: neural networks and random forests outperform the naive model, logistic regression, and the zero-inflated Poisson regression model. Also, based on Table II, it can be seen that while the naive model was fairly accurate, its accuracy was based on under-predicting accidents, as shown by its poor F1-score. Also, clustering (even in isolation) generally improved the F1-score and accuracy of the forecasting models (for each method, compare the two rows that denote no resampling). It can also be observed a similar trend with synthetic sampling, which even in isolation usually resulted in an improvement in accuracy as well as F1-score (for each method, see the set of rows that denote no clustering and compare the rows that show resampling). The efficacy of the combination of clustering and oversampling was somewhat unclear though. Typically, the combination slightly under-performs in comparison to using one of the two approaches. Three major takeaways can be drawn from this observation: first, synthetic sampling and clustering enabled forecasting in sparse datasets significantly more than approaches that do not use them. However, it is recommend that practitioners carefully evaluate each component of the proposed incident prediction pipeline on unseen data (test set) before deployment. Second, count-based models (zero-inflated Poisson regression) did not perform as well as binary classification models on sparse data. Third, it is important to note that while the resulting F1-scores might seem low in comparison to approaches on other data-driven problems, the improvement is significant in the context of extremely sparse and inherently random incidents like road accidents. The validity of this claim is shown later by simulating allocation and dispatch to accidents.

4.3 Feature Analysis

The proposed pipeline uses different features, such as congestion, precipitation, time of the day, etc. to predict the likelihood of accidents. However, not all the features are equally important. While some of the features may contribute more to the final predicted likelihood value, some of them may have less contribution or no contribution at all. While a comprehensive feature analysis requires a long-range of historical data, some of our fundamental analysis on features analysis is summarized below.

The goal is to investigate which features play a more important role in increasing the accident rate. To do so, the number of the incidents given that feature is counted and normalized to the frequency of that feature. The covariate “temperature” is used to explain the process. Then, this approach is used on all features and the most important features alongside their combinations with other features are reported.

Figure 4.2a shows the histogram of temperature values. The data is discretized into 4 bins (freezing, cold, mild, and hot) are shown using thin (and red) lines, Table II. Figure 4.2b shows the frequency of events for each category. Then, the number of accidents occurred during each one of the 4 temperature conditions is counted. In other words, the number of accidents given temperature condition is calculated, as shown in Figure 4.2c. It can be seen that most of the accidents occurred when the temperature was mild. However, this can be misleading since the frequency of mild temperature is also higher according to Figure 4.2b. Therefore, the parameter is normalized by dividing the frequency of each bin in Figure 4.2c by the frequency of the same bin in Figure 4.2d resulting in Figure 4.2d. It turns out accident rate on highways does not noticeably change based on temperature. However, the combination of temperature with other features might be important, which will be discussed later. In contrast, congestion has a sizeable impact on incidents. The other features are summarized in the Appendices section.

Table II: Categories of Features

#	feature	ranges	Tags
1	temperature	[min, 0, 10, 25, max]	['Freezing', 'Cold', 'Mild', 'Hot']
2	visibility	[min, 0.8, 3, max]	['Low', 'Fair', 'Clear']
3	precipitation	[min, 0.01, 1, max]	['No_Rain', 'Mild', 'Heavy']
4	Wind speed	[min, 3, 7, max]	['No_Wind', 'Mild', 'Windy']
5	congestion	[min, 0.1, 0.5, max]	['Light', 'Medium', 'Congested']
The other features are categorized using values and ranges. For example, below you can see how speed it categorized.			
6	speed	[0, 20, 40, 60, speed_max]	['[0-20)','[20-40)', '[40-60)','[60-max]'

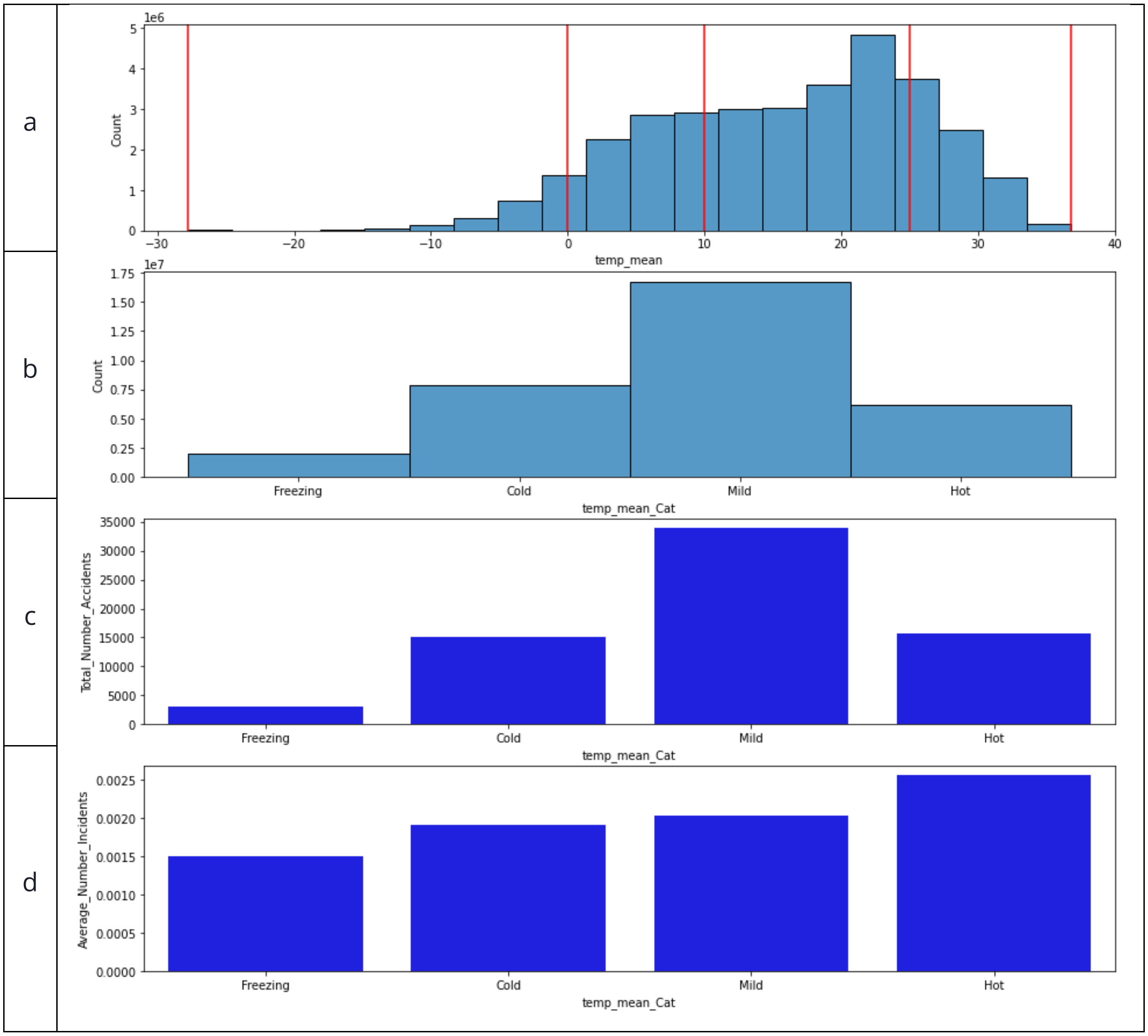


Figure 4.2 Feature analysis; estimation of importance of temperature in accident occurrence

Among the features, visibility, month, year, and wind speed are detected as not influential, while historical features are among the most important ones. Four important features were chosen, namely, temperature, precipitation, congestion, and weekend, to study the influence of their combination. A similar analysis is conducted but this time on a combination of features, summarized in Figure 4.3. It is observed that heavy precipitation combined with heavy (not light) congestion is an important determinant of accident occurrence, with a probability of occurrence almost seven times higher than the mean rate. While just freezing temperature slightly reduces the rate, its combination with precipitation increases the rate by almost a factor of two.

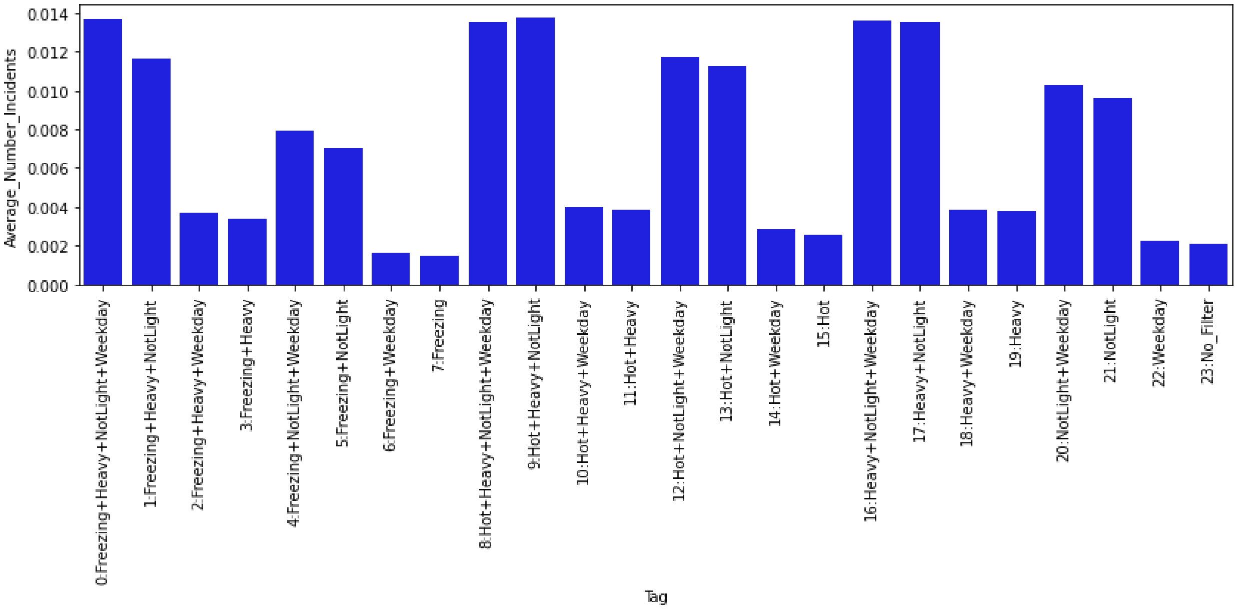


Figure 4.3 Feature analysis; estimation of importance of combination features in accident occurrence; the order of labels are temperature, precipitation, congestion, and weekend

4.4 Allocation and Dispatch

The final goal for this project is to enable our community partners save crucial response time to accidents. The entire combination of forecasting and dispatch was evaluated on 2190 temporal windows of 4 hours each. The hyper-parameter α and the number of available responders p were also varied. Two metrics were used to evaluate performance; the average distance traveled by the responders and the number of incidents that cannot be attended due to unavailability of responders. We presented the results in Figure 4.4 to Figure 4.6 as well as Table III to Table V.

An important observation is that our allocation approach, which adds a balancing term to the classical p-median problem, improved resource allocation in general; indeed, this was observed in general across the spectrum of forecasting model used and the number of available responders. The maximum improvement which was observed was a reduction of 3 km traveled by responders per incident (on average). This observation is particularly important for practitioners and first responders; while it is important to allocate resources in areas with (relatively) high historical rates of occurrence, assigning a small number of responders to cover large areas can be detrimental to the overall goal of reducing response times. Intuitively, the proposed approach penalizes additional burden on responders. However, it should also be noted that a large penalty (value of α) can result in increased response times. This is expected; as α grows, it discourages the geographic spread of responders. Figure 4.7 shows the influence of α on the performance of different models with a varying number of responders. The empirical results showed that $0.5 < \alpha < 1$ resulted in the optimal allocation of responders.

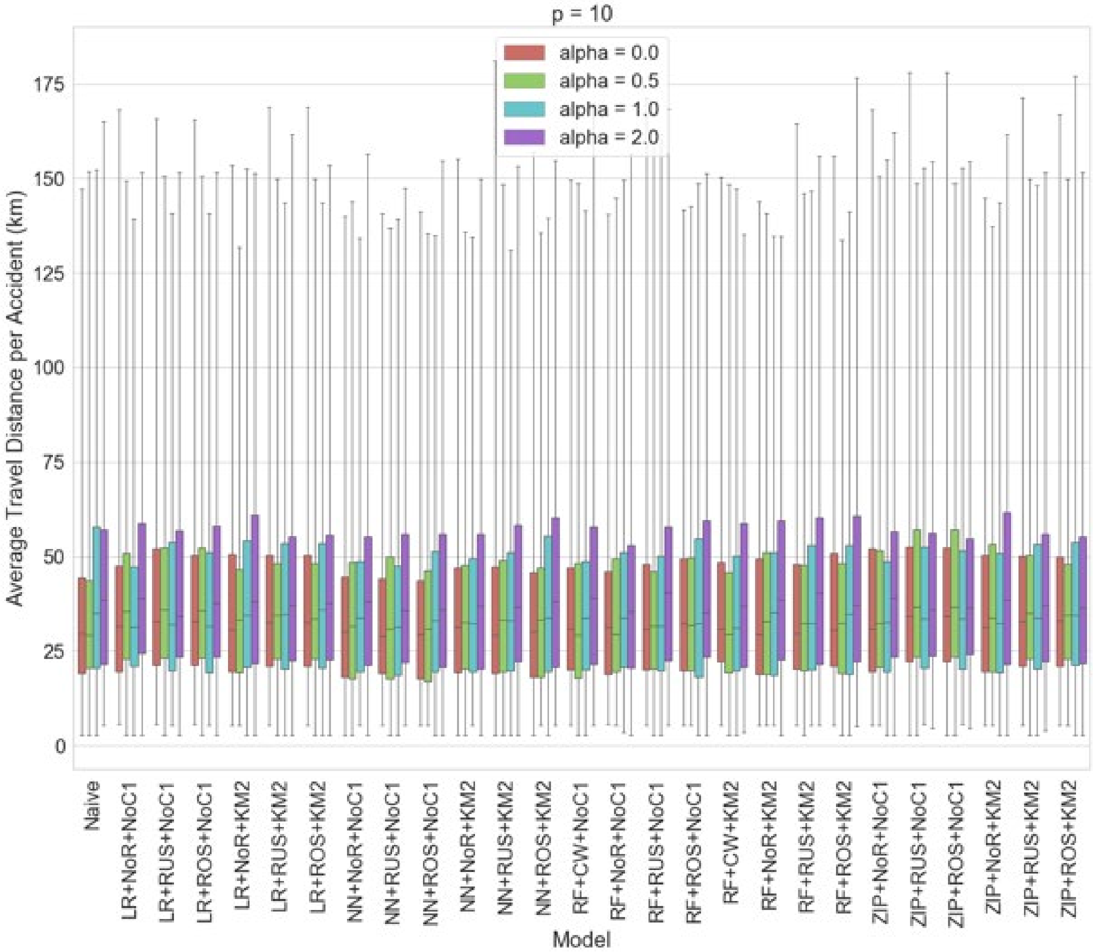


Figure 4.4 Total travel distance of responders per accident for 10; each model is evaluated for four different alpha values – from left to right 0, 0.5, 1, 2 (and also color coded)

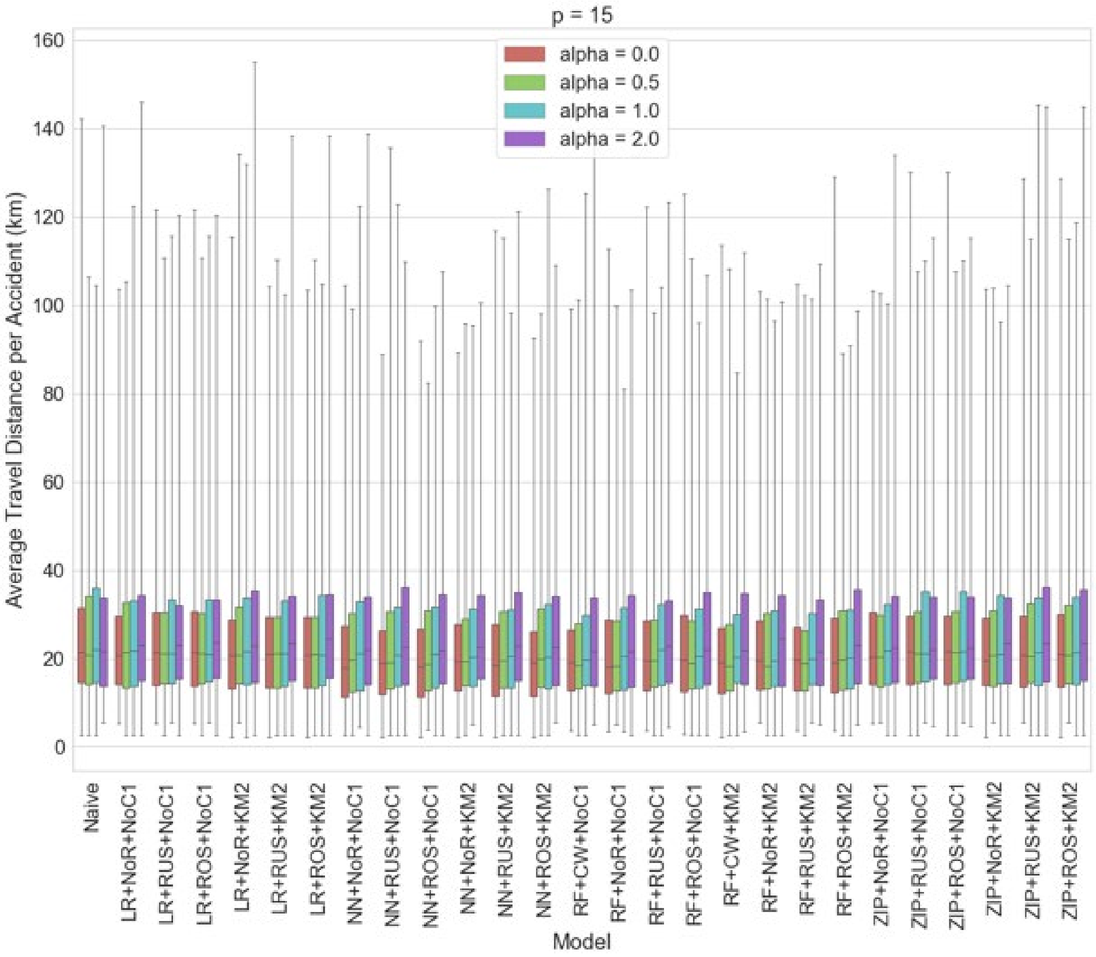


Figure 4.5 Total travel distance of responders per accident for 15; each model is evaluated for four different alpha values – from left to right 0, 0.5, 1, 2 (and also color coded)

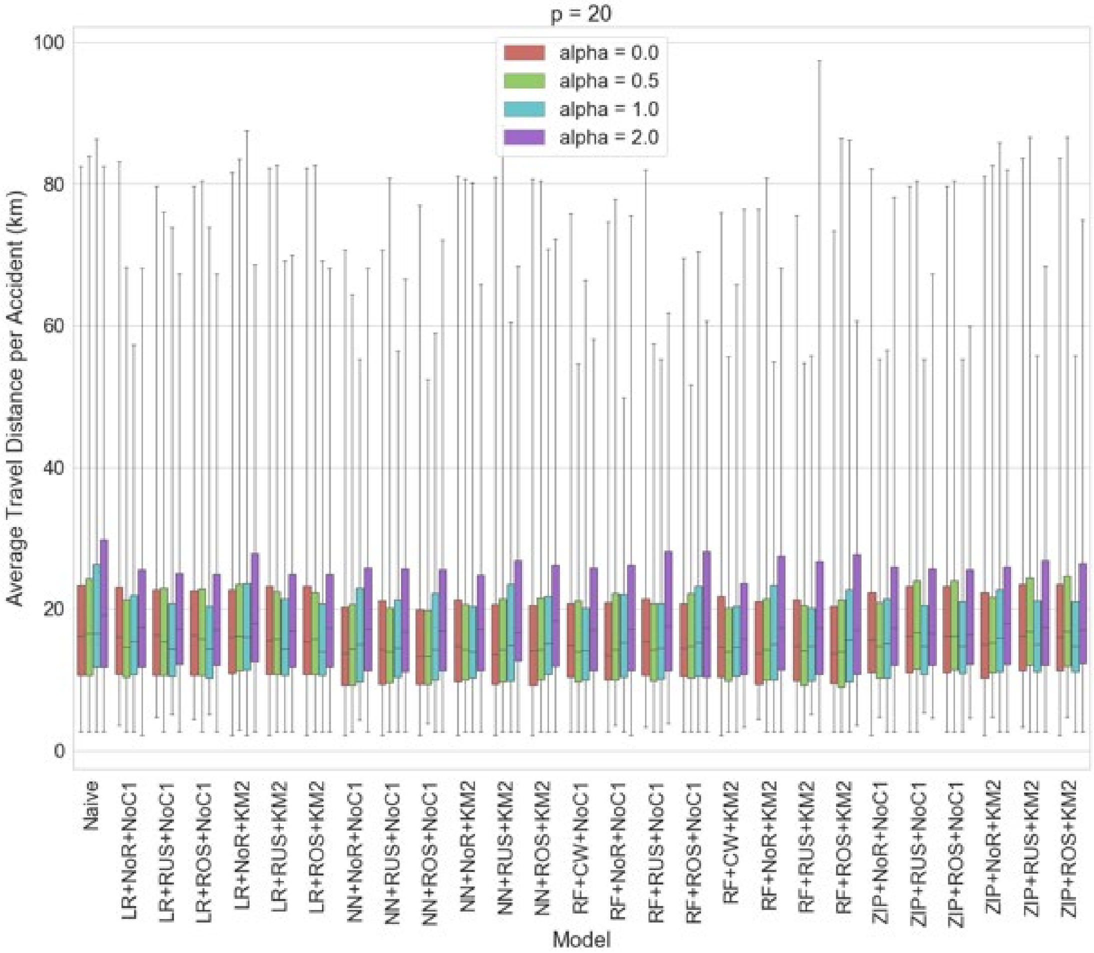


Figure 4.6 Total travel distance of responders per accident for 20; each model is evaluated for four different alpha values – from left to right 0, 0.5, 1, 2 (and also color coded)

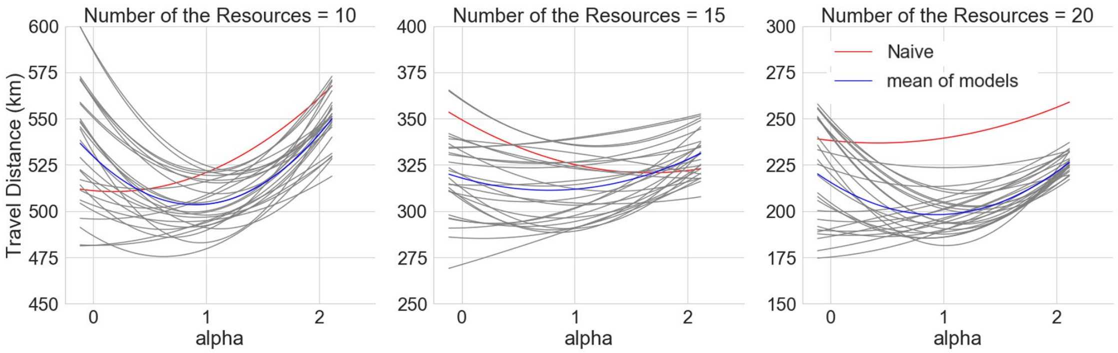


Figure 4.7 Effect the of the hyper parameterαin allocation model on the performance of different models (the grey lines denote each learned model)

It was also observed that the forecasting pipeline resulted in a noticeable improvement to response times (up to 19% for 20 available HELP trucks and up to 8% on average for multiple different numbers of available help trucks as it can be see in Figure 4.4 to Figure 4.6 and Table III) and reduced the total number of unattended accidents to (up to 75% and 50% for mean number and maximum number of unattended accidents during a 4-hour time window, respectively. Please see Table IV and Table V). In general, NN models provide the best results (and RF taking a close second). It is important to understand the importance of this reduction. Prior work reported that a saving of only ten minutes of response time can reduce deaths due to road accidents by 33% [101]. The major takeaways can be summarized based on the allocation and dispatch experiments. First, forecasting models that provide the highest accuracy might not be the best candidates for allocation. This observation shows the importance of using a metric that focuses on false negatives and false positives (like the F1-score) for sparse emergency incidents. Second, while traditional allocation models based on long-term (temporal) hotspots are widely used, accurate short-term forecasting models can result in significant reduction in response times to accidents. Finally, leveraging the structure of the problem to improve classical resource allocation formulations can aid emergency response in the field.

> our approach reduce response times up to **19%** for 20 available help trucks and up to **8%** on average for multiple different numbers of available help trucks.

Table III

Summary of performance evaluation metrics for each model in percentage (the performance in each column is color coded; green is the best and red is the worst)

Model	Clustering	Resampling	Name	Accuracy	Precision	Recall	F1	Pearson	Spearman
Naive			Naïve	**95.5**	3.8	4.2	4.0	**82.1**	60.8
LR	No cluster	No resampling	LR+NoR+NoC1	94.0	13.8	27.4	18.2	70.4	55.2
		RUS	LR+RUS+NoC1	93.0	12.8	32.3	18.3	63.1	54.7
		ROS	LR+ROS+NoC1	93.0	12.8	32.3	18.3	63.2	54.7
	clustering	No sample	LR+NoR+KM2	93.0	12.5	30.9	17.7	76.6	58.4
		RUS	LR+RUS+KM2	92.3	12.1	**34.4**	17.8	74.2	58.1
		ROS	LR+ROS+KM2	92.4	12.2	34.2	17.9	74.2	58.1
NN	No cluster	No resampling	NN+NoR+NoC1	94.9	19.2	32.8	24.0	71.7	58.5
		RUS	NN+RUS+NoC1	95.0	19.2	32.6	24.1	73.2	59.3
		ROS	NN+ROS+NoC1	94.9	19.1	32.8	23.9	69.3	54.7
	clustering	No sample	NN+NoR+KM2	95.0	19.0	31.6	23.7	75.6	58.9
		RUS	NN+RUS+KM2	94.7	18.4	32.7	23.3	73.1	54.6
		ROS	NN+ROS+KM2	94.7	18.3	33.1	23.3	74.5	55.4
Tree	No cluster	No resampling	RF+NoR+NoC1	95.0	19.0	31.8	23.6	78.7	63.4
		RUS	RF+RUS+NoC1	95.2	19.3	30.5	23.5	67.4	56.9
		ROS	RF+ROS+NoC1	95.3	18.6	27.6	22.1	79.2	**64.6**
		Class weights	RF+CW+NoC1	95.4	**20.6**	30.4	**24.4**	77.1	62.5
	clustering	No resampling	RF+NoR+KM2	95.1	18.9	30.5	23.2	79.8	62.3
		RUS	RF+RUS+KM2	95.0	19.4	32.5	24.2	73.8	57.6
		ROS	RF+ROS+KM2	95.1	18.3	28.7	22.2	80.1	63.6
		Class weights	RF+CW+NoC1	95.4	20.6	30.4	24.4	77.1	62.5
ZIP	No cluster	No resampling	ZIP+NoR+NoC1	94.4	14.6	26.8	18.9	74.0	58.0
		RUS	ZIP+RUS+NoC1	94.2	13.9	26.1	18.1	61.1	50.6
		ROS	ZIP+ROS+NoC1	94.2	13.9	26.7	18.2	61.2	50.6
	clustering	No resampling	ZIP+NoR+KM2	93.1	13.1	31.9	18.5	77.6	61.8
		RUS	ZIP+RUS+KM2	93.0	12.7	30.8	17.8	74.2	57.1
		ROS	ZIP+ROS+KM2	93.0	12.8	30.9	18.0	74.3	57.0

Table IV

Total travel distance of responders per accident in each4-h window

				p=10				*p*=15				*p*=20			
Model	clustering	Resampling	Name	*α*=0	*α*=0.5	*α*=1	*α*=0	*α*=0.5	*α*=1	*α*=0	*α*=0.5	*α*=1	*α*=0	*α*=0.5	*α*=1
Naive			Naïve	42.81	42.14	45.82	47.54	25.61	25.49	27.80	27.71	19.31	19.62	21.41	23.84
LR	No cluster	No resampling	LR+NoR+NoC1	43.06	42.67	42.08	46.83	25.11	24.57	25.83	27.57	18.59	17.30	17.52	21.21
		RUS	LR+RUS+NoC1	45.73	44.78	42.81	45.66	25.62	25.12	25.99	27.52	19.27	18.07	17.39	20.75
		ROS	LR+ROS+NoC1	45.69	44.72	42.67	45.65	25.64	25.10	25.93	27.55	19.28	18.03	17.35	20.74
	clustering	No sample	LR+NoR+KM2	42.76	42.56	43.35	47.27	24.31	24.65	26.02	28.38	18.54	18.73	19.01	21.90
		RUS	LR+RUS+KM2	44.62	42.96	43.06	46.76	24.75	24.62	25.76	27.97	18.68	18.72	17.72	20.84
		ROS	LR+ROS+KM2	44.72	42.89	43.04	46.70	24.75	24.56	25.88	27.90	18.70	18.72	17.74	20.75
NN	No cluster	No resampling	NN+NoR+NoC1	40.31	40.42	41.36	46.11	22.72	23.36	24.91	27.68	16.41	16.75	18.05	21.07
		RUS	NN+RUS+NoC1	40.05	40.45	41.30	46.58	22.47	23.42	24.86	28.43	16.18	16.86	18.18	21.53
		ROS	NN+ROS+NoC1	40.73	40.74	41.63	46.23	22.77	23.49	25.05	28.26	16.24	16.94	18.19	21.73
	clustering	No sample	NN+NoR+KM2	40.39	40.92	42.75	47.49	22.68	23.95	25.67	29.16	16.71	17.29	18.80	22.33
		RUS	NN+RUS+KM2	40.65	40.65	42.24	46.70	22.71	23.99	25.32	28.44	16.69	17.21	18.63	21.82
		ROS	NN+ROS+KM2	40.81	40.78	41.98	46.94	22.80	23.74	25.52	28.66	16.71	17.21	18.72	21.71
Tree	No cluster	No resampling	RF+NoR+NoC1	43.14	40.58	41.80	46.22	23.42	23.52	24.95	27.61	17.28	16.81	17.78	21.02
		RUS	RF+RUS+NoC1	41.77	40.29	42.43	47.10	23.26	23.63	25.54	28.83	16.71	17.38	18.70	21.46
		ROS	RF+ROS+NoC1	44.36	42.20	41.95	46.31	24.62	24.60	25.44	27.74	18.57	17.31	17.42	21.00
		Class weights	RF+CW+NoC1	42.34	41.13	42.35	47.09	23.33	23.94	25.49	28.82	16.72	17.35	18.44	21.76
	clustering	No resampling	RF+NoR+KM2	43.60	41.37	41.86	46.23	23.71	23.77	25.05	27.57	17.56	16.99	17.84	20.90
		RUS	RF+RUS+KM2	42.00	41.18	42.77	48.10	23.29	24.12	25.79	29.03	16.79	17.56	18.85	21.89
		ROS	RF+ROS+KM2	43.20	41.13	42.08	46.25	23.86	23.89	25.24	27.53	17.55	16.98	17.95	21.03
		Class weights	RF+CW+NoC1	42.54	40.95	42.27	47.59	23.65	24.29	25.56	28.79	17.04	17.49	18.81	21.55
ZIP	No cluster	No resampling	ZIP+NoR+NoC1	42.37	42.06	41.53	46.10	24.29	24.17	25.12	27.76	18.00	17.07	17.54	21.21
		RUS	ZIP+RUS+NoC1	47.17	46.73	43.20	46.01	26.45	25.74	26.46	27.57	19.59	19.13	17.78	20.52
		ROS	ZIP+ROS+NoC1	47.21	46.74	43.14	46.10	26.43	25.74	26.54	27.61	19.55	19.12	17.79	20.59
	clustering	No resampling	ZIP+NoR+KM2	42.22	42.14	42.56	46.52	23.88	24.35	25.89	28.14	17.92	18.21	18.68	21.93
		RUS	ZIP+RUS+KM2	45.95	43.50	43.97	46.66	25.37	25.20	26.41	28.31	19.22	19.44	17.76	21.20
		ROS	ZIP+ROS+KM2	46.02	43.47	44.00	46.81	25.39	25.22	26.38	28.34	19.20	19.51	17.71	21.19

Table V

Average number accidents in each 4-hwindow that responders are not able to immediately respond because all responders are busy (the performance in each column is color coded; green is the best and red is the worst)

				p=10				*p*=15				*p*=20			
Model	clustering	Resampling	Name	*α*=0	*α*=0.5	*α*=1	*α*=2	*α*=0	*α*=0.5	*α*=1	*α*=2	*α*=0	*α*=0.5	*α*=1	*α*=2
Naive			Naïve	0.59	0.55	0.51	0.53	0.06	0.05	0.05	0.05	0.01	0.01	0.01	0.01
LR	No cluster	No resampling	LR+NoR+NoC1	0.57	0.49	0.49	0.49	0.06	0.05	0.04	0.04	0.00	0.00	0.00	0.00
		RUS	LR+RUS+NoC1	0.61	0.54	0.52	0.51	0.06	0.05	0.05	0.04	0.01	0.00	0.00	0.00
		ROS	LR+ROS+NoC1	0.62	0.54	0.52	0.51	0.06	0.05	0.05	0.04	0.01	0.00	0.00	0.00
	clustering	No sample	LR+NoR+KM2	0.56	0.49	0.48	0.51	0.06	0.05	0.05	0.05	0.00	0.00	0.00	0.00
		RUS	LR+RUS+KM2	0.60	0.52	0.49	0.52	0.06	0.05	0.05	0.05	0.00	0.00	0.00	0.00
		ROS	LR+ROS+KM2	0.60	0.51	0.50	0.52	0.06	0.05	0.05	0.05	0.00	0.00	0.00	0.00
NN	No cluster	No resampling	NN+NoR+NoC1	0.51	0.47	0.48	0.50	0.04	0.04	0.04	0.05	0.00	0.00	0.00	0.00
		RUS	NN+RUS+NoC1	0.49	0.46	0.46	0.50	0.04	0.04	0.04	0.04	0.00	0.00	0.00	0.00
		ROS	NN+ROS+NoC1	0.51	0.47	0.48	0.49	0.05	0.04	0.05	0.04	0.00	0.00	0.00	0.00
	clustering	No sample	NN+NoR+KM2	0.50	0.47	0.48	0.51	0.04	0.04	0.04	0.04	0.00	0.00	0.00	0.00
		RUS	NN+RUS+KM2	0.51	0.47	0.47	0.50	0.05	0.05	0.04	0.05	0.00	0.00	0.00	0.00
		ROS	NN+ROS+KM2	0.51	0.47	0.47	0.50	0.05	0.04	0.05	0.04	0.00	0.00	0.00	0.00
Tree	No cluster	No resampling	RF+NoR+NoC1	0.57	0.48	0.49	0.50	0.05	0.04	0.04	0.04	0.00	0.00	0.00	0.00
		RUS	RF+RUS+NoC1	0.53	0.47	0.47	0.50	0.05	0.05	0.04	0.04	0.00	0.00	0.00	0.00
		ROS	RF+ROS+NoC1	0.61	0.52	0.50	0.50	0.06	0.05	0.05	0.04	0.01	0.00	0.00	0.01
		Class weights	RF+CW+NoC1	0.54	0.48	0.46	0.48	0.05	0.04	0.04	0.04	0.00	0.00	0.00	0.00
	clustering	No resampling	RF+NoR+KM2	0.58	0.50	0.48	0.51	0.05	0.04	0.05	0.04	0.01	0.00	0.00	0.00
		RUS	RF+RUS+KM2	0.53	0.47	0.47	0.50	0.05	0.05	0.04	0.04	0.00	0.00	0.00	0.00
		ROS	RF+ROS+KM2	0.58	0.50	0.48	0.50	0.05	0.05	0.05	0.04	0.00	0.00	0.00	0.00
		Class weights	RF+CW+NoC1	0.55	0.48	0.47	0.49	0.05	0.05	0.04	0.04	0.00	0.00	0.00	0.00
ZIP	No cluster	No resampling	ZIP+NoR+NoC1	0.54	0.49	0.46	0.48	0.06	0.04	0.04	0.04	0.00	0.00	0.00	0.00
		RUS	ZIP+RUS+NoC1	0.64	0.56	0.54	0.52	0.06	0.06	0.05	0.05	0.01	0.00	0.00	0.00
		ROS	ZIP+ROS+NoC1	0.64	0.56	0.54	0.52	0.06	0.06	0.05	0.05	0.01	0.00	0.00	0.00
	clustering	No resampling	ZIP+NoR+KM2	0.54	0.49	0.47	0.49	0.06	0.04	0.05	0.05	0.00	0.00	0.00	0.00
		RUS	ZIP+RUS+KM2	0.62	0.53	0.51	0.51	0.05	0.05	0.05	0.05	0.01	0.00	0.00	0.00
		ROS	ZIP+ROS+KM2	0.63	0.53	0.52	0.51	0.05	0.05	0.05	0.05	0.01	0.00	0.00	0.00

Table VI

Maximum number accidents in each 4-hwindow that responders are not able to immediately respond because all responders are busy (the performance in each column is color coded; green is the best and red is the worst)

				p=10				p=15				p=20			
Model	clustering	Resampling	Name	α=0	α=0.5	α=1	α=2	α=0	α=0.5	α=1	α=2	α=0	α=0.5	α=1	α=2
Naive			Naïve	31.00	34.00	33.00	34.00	22.00	22.00	22.00	17.00	9.00	10.00	10.00	6.00
LR	No cluster	No resampling	LR+NoR+NoC1	32.00	31.00	30.00	31.00	21.00	21.00	15.00	16.00	6.00	5.00	4.00	3.00
		RUS	LR+RUS+NoC1	31.00	31.00	33.00	32.00	19.00	18.00	21.00	14.00	8.00	6.00	6.00	5.00
		ROS	LR+ROS+NoC1	31.00	31.00	33.00	32.00	19.00	18.00	21.00	14.00	8.00	6.00	6.00	5.00
	clustering	No sample	LR+NoR+KM2	33.00	31.00	34.00	31.00	22.00	21.00	19.00	20.00	7.00	8.00	6.00	4.00
		RUS	LR+RUS+KM2	31.00	32.00	32.00	33.00	23.00	21.00	19.00	16.00	7.00	5.00	8.00	5.00
		ROS	LR+ROS+KM2	31.00	31.00	32.00	32.00	23.00	23.00	19.00	15.00	7.00	6.00	8.00	6.00
NN	No cluster	No resampling	NN+NoR+NoC1	31.00	32.00	35.00	34.00	19.00	16.00	18.00	18.00	6.00	4.00	7.00	5.00
		RUS	NN+RUS+NoC1	31.00	31.00	34.00	34.00	19.00	20.00	17.00	16.00	5.00	6.00	6.00	4.00
		ROS	NN+ROS+NoC1	31.00	34.00	34.00	34.00	21.00	19.00	19.00	17.00	6.00	6.00	6.00	4.00
	clustering	No sample	NN+NoR+KM2	30.00	35.00	34.00	34.00	20.00	19.00	19.00	15.00	7.00	6.00	6.00	4.00
		RUS	NN+RUS+KM2	30.00	32.00	35.00	34.00	20.00	19.00	18.00	20.00	8.00	6.00	6.00	7.00
		ROS	NN+ROS+KM2	32.00	32.00	32.00	34.00	19.00	17.00	18.00	15.00	7.00	6.00	7.00	5.00
Tree	No cluster	No resampling	RF+NoR+NoC1	33.00	31.00	32.00	34.00	20.00	20.00	18.00	16.00	7.00	5.00	5.00	7.00
		RUS	RF+RUS+NoC1	31.00	32.00	34.00	35.00	21.00	21.00	19.00	17.00	6.00	6.00	5.00	6.00
		ROS	RF+ROS+NoC1	33.00	31.00	32.00	32.00	22.00	19.00	23.00	19.00	8.00	6.00	8.00	9.00
		Class weights	RF+CW+NoC1	31.00	31.00	31.00	33.00	22.00	18.00	17.00	15.00	8.00	9.00	6.00	5.00
	clustering	No resampling	RF+NoR+KM2	33.00	30.00	31.00	34.00	22.00	20.00	19.00	14.00	9.00	4.00	5.00	5.00
		RUS	RF+RUS+KM2	32.00	31.00	34.00	35.00	21.00	21.00	13.00	18.00	5.00	6.00	5.00	7.00
		ROS	RF+ROS+KM2	31.00	31.00	31.00	32.00	19.00	20.00	18.00	15.00	6.00	5.00	7.00	6.00
		Class weights	RF+CW+NoC1	32.00	31.00	33.00	35.00	22.00	21.00	21.00	17.00	5.00	6.00	8.00	8.00
ZIP	No cluster	No resampling	ZIP+NoR+NoC1	32.00	33.00	31.00	33.00	21.00	19.00	20.00	16.00	6.00	5.00	7.00	3.00
		RUS	ZIP+RUS+NoC1	31.00	32.00	33.00	32.00	19.00	21.00	21.00	18.00	8.00	7.00	6.00	7.00
		ROS	ZIP+ROS+NoC1	31.00	32.00	33.00	34.00	19.00	21.00	21.00	20.00	8.00	7.00	6.00	7.00
	clustering	No resampling	ZIP+NoR+KM2	33.00	32.00	33.00	34.00	21.00	20.00	18.00	17.00	7.00	5.00	6.00	3.00
		RUS	ZIP+RUS+KM2	31.00	32.00	34.00	34.00	23.00	21.00	19.00	17.00	8.00	7.00	6.00	7.00
		ROS	ZIP+ROS+KM2	31.00	32.00	32.00	34.00	23.00	18.00	20.00	17.00	8.00	7.00	6.00	7.00

4.5 Implementation

The implementation is open-source and can be used by other organizations that seek to optimize emergency response. The modularized designed pipeline is shown in Figure 4.8. The whole pipeline is dockerized[5] and will be provided to TDOT for operation. It has four main modules as follows:

Data Collection and Merging Datasets: This module collects data from various resources including Inrix for traffic, Inrix and Google Cloud Elevation API and Inrix for roadway segments, TDOT for historical accident data, and Weatherbit for weather information. Then, it combines and aggregates them for the intended time and spatial resolution. Most of this pipeline uses cloud resources.

Data Preparation: This module prepares the datasets required for our machine learning engine. Based on the type of clustering, number of clusters, resampling scenario, and resampling ratio, this module creates 3 datasets (train, validation, test).

Training Models: This module is the heart of the proposed prediction engine. It uses various machine learning model to predict the spatial temporal likelihood of incidents. It uses performance metrics (accuracy, precision, recall, F1score, Pearson and Spearman correlation, correctness, etc.) to evaluate the trained models.

Prediction Module: After training various models, the best one is selected to spatially and temporarily predict the likelihood of incidents. An initial best trained model has been provided in the toolchain, but the model can be updated using the training workflow also included in the toolchain. To help with prediction, a submodule is included to collect the required information for the future. For example, Weatherbit API is used, which facilitates the collection weather data up to 5 days in the future.

Simulation Module: Based on the predicted likelihood of incidents, Simulation module uses the modified P-median approach to distribute limited resources (number of ambulances defined by user) strategically on the map. Then it uses the incident time and locations to run a real-time scenario to evaluate the performance of the prediction by calculating the response time.

Each of the modules can be deployed using Docker, the provided docker image, defined config file, location of input files, and the location of output files. The docker commands are summerized in Table VI. For futher details, please refer to Docker section in the Appendices.

[5] https://www.docker.com/

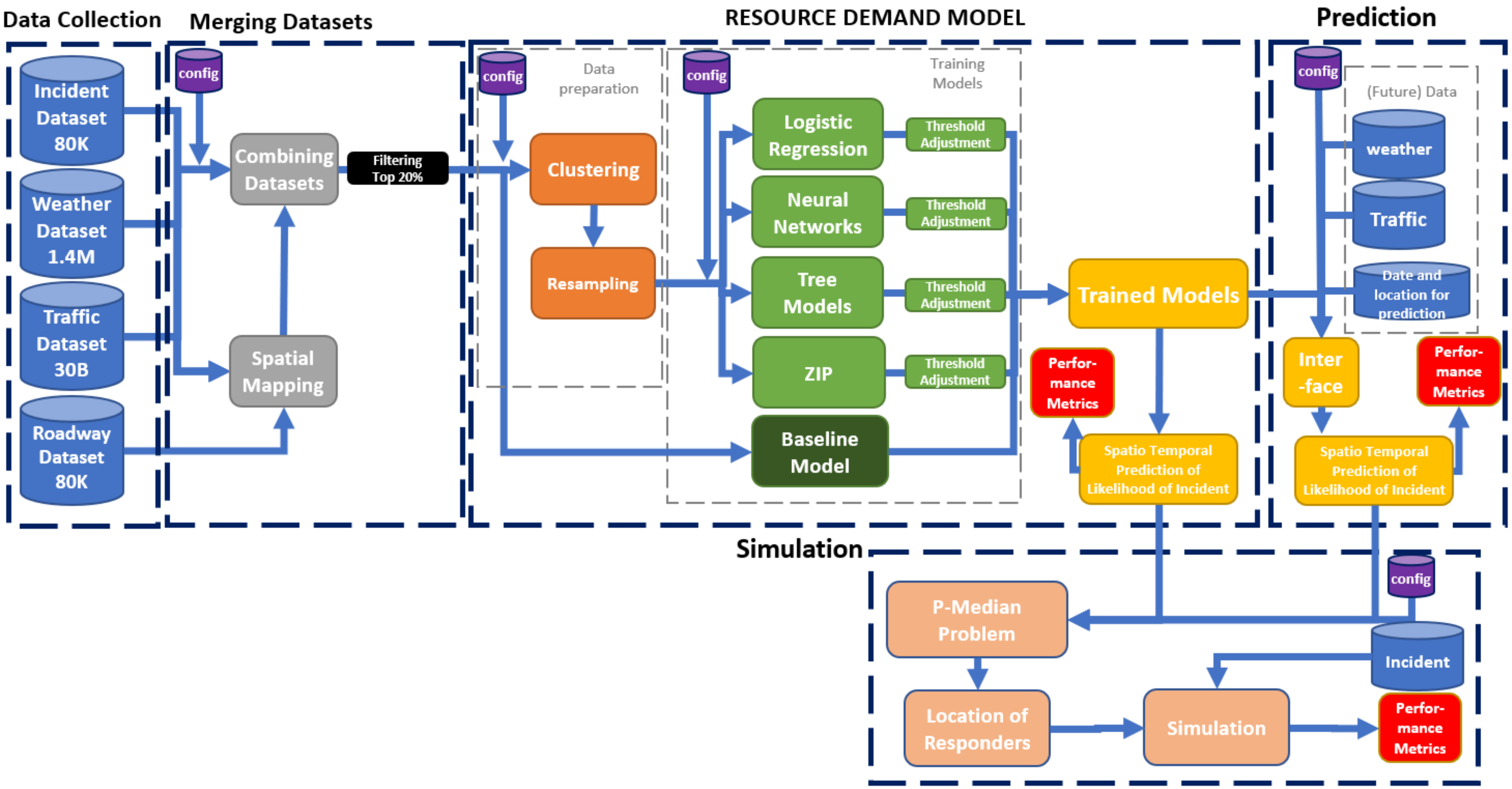
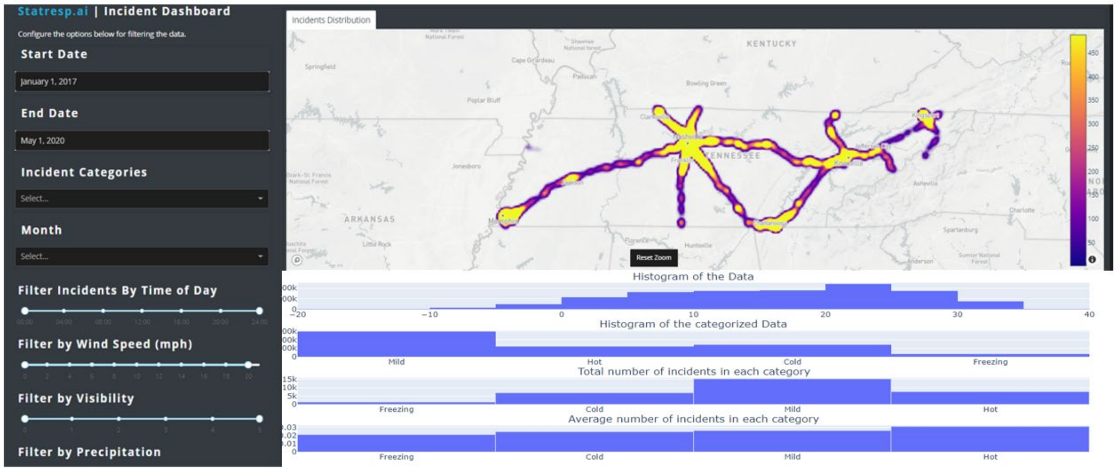


Figure 4.8 Overview of the designed modular pipeline

Table VII
Summary of the docker commands

Serial	Modeule	Function	Defined flags	Mounting folders	Final Command
1	*Data Preparation*	run_dataprep .py	-c /app/etc/config.conf -i /app/data -o /app/output1	-v "%cd%/etc":/app/etc -v "%cd%/data":/app/data -v "%cd%/output1":/app/output1	docker run -it -v "%cd%/data":/app/data -v "%cd%/etc":/app/etc -v "%cd%/output1":/app/output1 prediction_engine_1 run_dataprep.py -i /app/data -c /app/etc/config.conf -o /app/output1
2	*Training Model*	run_training .py			docker run -it -v "%cd%/data":/app/data -v "%cd%/etc":/app/etc -v "%cd%/output1":/app/output1 prediction_engine_1 run_prediction.py -i /app/data -c /app/etc/config.conf -o /app/output1
3	*Prediction*	run_prediciton .py			docker run -it -v "%cd%/data":/app/data -v "%cd%/etc":/app/etc -v "%cd%/output1":/app/output1 prediction_engine_1 run_prediction.py -i /app/data -c /app/etc/config.conf -o /app/output1
4	*Simulation*	run_simulation .py			docker run -it -v "%cd%/data":/app/data -v "%cd%/etc":/app/etc -v "%cd%/output1":/app/output1 prediction_engine_1 run_simulation.py -i /app/data -c /app/etc/config.conf -o /app/output1

Furthermore, a user interface (dashboard) has been designed to improve the user experience of the pipeline. Two separate dashboards were also designed; one for visualizing the features and historical accidents and another one for visualizing the prediction of likelihood of incident occurrence, location of responders relative to incidents, and the performance of the prediction based on assorted metrics. Figure 4.9 and Figure 4.10 are the drafts of the view of the dashboard of historical mode and prediction and evaluation mode, respectively. The dashboards are available using https://mystic-impulse-228617.ue.r.appspot.com/.



Statresp.ai | Incident Dashboard

Configure the options below for filtering the data.

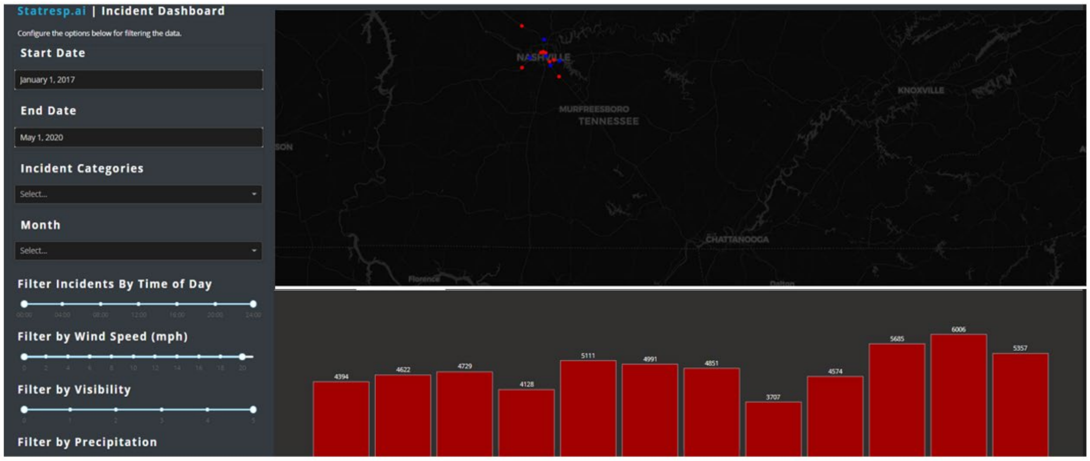
Start Date: January 1, 2017

End Date: May 1, 2020

Incident Categories: Select...

Month: Select...

Figure 4.9 Dashboard in historical mode



Statresp.ai | Incident Dashboard

Configure the options below for filtering the data.

Start Date

January 1, 2017

End Date

May 1, 2020

Incident Categories

Select...

Month

Select...

Filter Incidents By Time of Day

Filter by Wind Speed (mph)

Filter by Visibility

Filter by Precipitation

Figure 4.10 Dashboard in prediction and evaluation mode

4.6 Discussion

Emergency response to incidents like road accidents is a major concern for first responders. Standard approaches to predict road accidents rarely scale to large geographic areas due to extremely high sparsity in data and difficulties in gathering data. In collaboration with TDOT, a framework is presented for forecasting extremely sparse spatial and temporal incidents like road accidents. We show how our approach for forecasting, based on a combination of non-spatial clustering, synthetic resampling, and learning from multiple data sources, outperforms forecasting methods used in the field. A novel modification is also presented to a classical formulation for resource allocation. Through extensive simulations, it is shown how our pipeline results in significant reduction in response times to emergency incidents and unattended number of incidents due to unavailablity of the resources.

The modularized designed of the pipeline enables the replacement of each module or in the future. Even though the proposed approach showed promising results, the geographic spread of the model can be improved by collecting more accident data. With the availability of more incidents in a broader spatial area and longer time range, more advanced machine learning approaches such as model stacking technique can also be employed. However, if the available data is from another environment (for example historical accident data from another state such as California), transfer learning techniques can be used to reduce the detrimental influence of insufficient data. It is worth investigating the role more advanced clustering methods. Moreover, incorporating spatial correlation by using graph theory and modeling segments as nodes of a graph and connecting the neighboring segments (nodes). While this research was heavily focused on prediction, combining prediction with detection, in particular by leveraging the crowd-sourced data platforms such as Waze, can improve the results. The aforementioned items will be investigated in the next phase of the research. Currently, our model is trained for drastic changes in the traffic flow by external factors such as COVID-19 pandemic. By the availability of the new incident dataset containing the incidents after April/2020, the influence COVID-19 pandemic in accident patterns can also be investigated.

Chapter 5 Conclusion

Emergency response to incidents like road accidents is a major concern for first responders. Standard approaches to predict road accidents rarely scale to large geographic areas due to extremely high sparsity in data and difficulties in gathering data. In collaboration with TDOT, a framework for forecasting extremely sparse spatio-temporal incidents like road accidents is presented. We show how the proposed approach for forecasting, based on a combination of non-spatial clustering, synthetic resampling, and learning from multiple data sources, outperforms forecasting methods used in the field. A novel modification to a classical formulation for resource allocation is also presented. Through extensive simulations, we show how our pipeline results in a significant improvement. This open-source implementation can be used by other organizations looking for emergency response optimization.

First all the accessible information was evaluated from various resources that can be useful in incident prediction and designed an efficient pipeline to collect, clean, and combine them, which can be used in future for any other similar research. The pipeline facilitates the process of keeping the final data updated. A pipeline was also designed to forecast the spatial and temporal dynamics of accident occurrence, even under sparse conditions by combining assorted machine learning techniques. Using a battery of metrics, it was shown that our model outperforms the current conventional approach in the field. Due to modularity of the code, it is easy to replace each module in the code with a more advanced or more efficient methods in the future. The proposed incident prediction pipeline can be used for strategic allocation of the resources (HELP trucks) leading to faster response time, increasing the safety of the highways, and reducing direct and indirect costs for the state.

References

[1] A. Mukhopadhyay *et al.*, "A Review of Incident Prediction, Resource Allocation, and Dispatch Models for Emergency Management." 2021.

[2] "Disaster Sequence of Events." https://training.fema.gov/emiweb/downloads/is208sdmunit3.pdf.

[3] A. Mukhopadhyay, C. Zhang, Y. Vorobeychik, M. Tambe, K. Pence, and P. Speer, "Optimal Allocation of Police Patrol Resources Using a Continuous-Time Crime Model," 2016, pp. 139–158.

[4] A. Mukhopadhyay, Y. Vorobeychik, A. Dubey, and G. Biswas, "Prioritized Allocation of Emergency Responders Based on a Continuous-Time Incident Prediction Model," in *16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2017, pp. 168–177.

[5] A. Mukhopadhyay, G. Pettet, C. Samal, A. Dubey, and Y. Vorobeychik, "An Online Decision-Theoretic Pipeline for Responder Dispatch," in *10th ACM/IEEE International Conference on Cyber-Physical Systems, (ICCPS)*, 2019, pp. 185–196, doi: 10.1145/3302509.3311055.

[6] G. Pettet, A. Mukhopadhyay, M. Kochenderfer, Y. Vorobeychik, and A. Dubey, "On Algorithmic Decision Procedures in Emergency Response Systems in Smart and Connected Communities," Jan. 2020, Accessed: Feb. 02, 2020. [Online]. Available: http://arxiv.org/abs/2001.07362.

[7] G. Pettet, A. Mukhopadhyay, M. Kochenderfer, and A. Dubey, "Hierarchical Planning for Resource Allocation in Emergency Response Systems," *arXiv Prepr. arXiv2012.13300*, 2020.

[8] J. A. Deacon, C. V Zegeer, and R. C. Deen, "Identification of Hazardous Rural Highway Locations," *Transp. Res. Rec.*, vol. 543, 1974.

[9] M. Deublein, M. Schubert, B. T. Adey, J. Köhler, and M. H. Faber, "Prediction of road accidents: A Bayesian hierarchical approach," *Accid. Anal. {\&} Prev.*, vol. 51, pp. 274–291, 2013, doi: https://doi.org/10.1016/j.aap.2012.11.019.

[10] M. A. Quddus, "Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data," *Accid. Anal. Prev.*, vol. 40, no. 4, pp. 1486–1497, Jul. 2008, doi: 10.1016/j.aap.2008.03.009.

[11] D. Akin, "Analysis of Highway Crash Data by Negative Binomial and Poisson Regression Models," 2011, doi: 10.13140/2.1.4113.0567.

[12] F. de Guevara, S. P. Washington, and J. Oh, "Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona," *Transp. Res. Rec.*, vol. 1897, no. 1, pp. 191–199, 2004.

[13] W. Ackaah and M. Salifu, "Crash Prediction Model for Two-Lane Rural Highways in the Ashanti Region of Ghana," *IATSS Res.*, vol. 35, no. 1, pp. 34–40, 2011.

[14] X. Qin, J. N. Ivan, and N. Ravishanker, "Selecting exposure measures in crash rate prediction for two-lane highway segments," *Accid. Anal. \& Prev.*, vol. 36, no. 2, pp. 183–191, 2004.

[15] D. Lord, S. Washington, and J. N. Ivan, "Further notes on the application of zero-inflated models in highway safety," *Accid. Anal. \& Prev.*, vol. 39, no. 1, pp. 53–57, 2007.

[16] H. Huang and H. C. Chin, "Modeling road traffic crashes with zero-inflation and site-specific

random effects," *Stat. Methods \& Appl.*, vol. 19, no. 3, pp. 445–462, 2010.

[17] D. Lord, S. P. Washington, and J. N. Ivan, "Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory," *Accid. Anal. Prev.*, vol. 37, no. 1, pp. 35–46, 2005, doi: https://doi.org/10.1016/j.aap.2004.02.004.

[18] A. Pande and M. Abdel-Aty, "Assessment of freeway traffic parameters leading to lane-change related collisions," *Accid. Anal. Prev.*, vol. 38, no. 5, pp. 936–948, Sep. 2006, doi: 10.1016/J.AAP.2006.03.004.

[19] H. T. Abdelwahab and M. A. Abdel-Aty, "Artificial Neural Networks and Logit Models for Traffic Safety Analysis of Toll Plazas," *Transp. Res. Rec.*, vol. 1784, no. 1, pp. 115–125, Jan. 2002, doi: 10.3141/1784-15.

[20] L.-Y. Chang, "Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network," *Saf. Sci.*, vol. 43, no. 8, pp. 541–557, 2005, doi: 10.1016/j.ssci.2005.04.004.

[21] C. Riviere, P. Lauret, J.-F. Ramsamy Manicom, and Y. Page, "A Bayesian neural network approach to estimating the energy equivalent speed," *Accid. Anal. Prev.*, vol. 38, no. 2, pp. 248–259, 2006, doi: 10.1016/j.aap.2005.08.008.

[22] L. Zhu, F. Guo, R. Krishnan, and J. W. Polak, "The use of convolutional neural networks for traffic incident detection at a network level," 2018.

[23] J. Bao, P. Liu, and S. V Ukkusuri, "A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data," *Accid. Anal. {\&} Prev.*, vol. 122, pp. 239–254, 2019, doi: 10.1016/J.AAP.2018.10.015.

[24] Y. Zhang and Y. Xie, "Forecasting of Short-Term Freeway Volume with v-Support Vector Machines," *Transp. Res. Rec.*, vol. 2024, no. 1, pp. 92–99, 2007, doi: 10.3141/2024-11.

[25] X. Li, D. Lord, Y. Zhang, and Y. Xie, "Predicting motor vehicle crashes using Support Vector Machine models," *Accid. Anal. Prev.*, vol. 40, no. 4, pp. 1611–1618, 2008, doi: 10.1016/j.aap.2008.04.010.

[26] R. Yu and M. Abdel-Aty, "Utilizing support vector machine in real-time crash risk evaluation," *Accid. Anal. {\&} Prev.*, vol. 51, pp. 252–259, 2013, doi: 10.1016/J.AAP.2012.11.027.

[27] J. Frantzeskakis, V. Assimakopoulos, and G. Kindinis, "Interurban Accident Prediction by Administrative Area Application in Greece," *Inst. Transp. Eng. J.*, vol. 64, no. 1, pp. 35–42, 1994.

[28] A. Osei-Asamoah and N. E. Lownes, "Complex Network Method of Evaluating Resilience in Surface Transportation Networks," *Transp. Res. Rec.*, vol. 2467, no. 1, pp. 120–128, Jan. 2014, doi: 10.3141/2467-13.

[29] P. P. Jovanis and H.-L. Chang, "Modeling the Relationship of Accidents to Miles Traveled," *Transp. Res. Rec.*, vol. 1068, pp. 42–51, 1986.

[30] S.-P. Miaou and H. Lum, "Modeling vehicle accidents and highway geometric design relationships," *Accid. Anal. \& Prev.*, vol. 25, no. 6, pp. 689–709, 1993.

[31] S. C. Joshua and N. J. Garber, "Estimating Truck Accident Rate and Involvements using Linear and Poisson Regression Models," *Transp. Plan. Technol.*, vol. 15, no. 1, pp. 41–58, 1990.

[32] V. Niveditha, A. Ramesh, and M. Kumar, "Development of Models for Crash Prediction and

Collision Estimation - A Case Study for Hyderabad City," *Int. J. Transp. Eng.*, vol. 3, no. 2, pp. 143–150, 2015.

[33] H. Rakha, M. Arafeh, A. G. Abdel-Salam, F. Guo, and A. M. Flintsch, "Linear Regression Crash Prediction Models: Issues and Proposed Solutions," *Effic. Transp. Pavement Syst. Charact. Mech. Simul. Model.*, pp. 241–256, 2010.

[34] T. Sayed and F. Rodriguez, "Accident Prediction Models for Urban Unsignalized Intersections in British Columbia," *Transp. Res. Rec. J. Transp. Res. Board*, no. 1665, pp. 93–99, 1999.

[35] J. A. Bonneson and P. T. McCoy, "Estimation of Safety at Two-Way Stop-Controlled Intersections on Rural Highways," *Transp. Res. Rec.*, no. 1401, pp. 83–89, 1993.

[36] M. J. Maher and I. Summersgill, "A Comprehensive Methodology for the Fitting of Predictive Accident Models," *Accid. Anal. \& Prev.*, vol. 28, no. 3, pp. 281–296, 1996.

[37] Z. Ye, Y. Xu, and D. Lord, "Crash Data Modeling with a Generalized Estimator," *Accid. Anal. \& Prev.*, vol. 117, pp. 340–345, 2018.

[38] C. Caliendo, M. Guida, and A. Parisi, "A Crash-Prediction Model for Multilane Roads," *Accid. Anal. \& Prev.*, vol. 39, no. 4, pp. 657–670, 2007.

[39] S. Dissanayake and I. Ratnayake, "Statistical Modelling of Crash Frequency on Rural Freeways and Two-Lane Highways using Negative Binomial Distribution," *Adv. Transp. Stud.*, vol. 9, pp. 81–96, 2006.

[40] D. Lord and L. F. Miranda-Moreno, "Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter of Poisson-Gamma Models for Modeling Motor Vehicle Crashes: A Bayesian Perspective," *Saf. Sci.*, vol. 46, no. 5, pp. 751–770, 2008.

[41] J. Aguero-Valverde and P. P. Jovanis, "Analysis of road crash frequency with spatial models," *Transp. Res. Rec.*, vol. 2061, no. 1, pp. 55–63, 2008.

[42] M. Shirazi and D. Lord, "Characteristics-based heuristics to select a logical distribution between the Poisson-gamma and the Poisson-lognormal for crash data modelling," *Transp. A Transp. Sci.*, vol. 15, no. 2, pp. 1791–1803, Nov. 2019, doi: 10.1080/23249935.2019.1640313.

[43] E. S. Park and D. Lord, "Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2019, no. 1, pp. 1–6, Jan. 2007, doi: 10.3141/2019-01.

[44] J. Ma, K. M. Kockelman, and P. Damien, "A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods," *Accid. Anal. Prev.*, vol. 40, no. 3, pp. 964–975, May 2008, doi: 10.1016/J.AAP.2007.11.002.

[45] J. Aguero-Valverde, "Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates," *Accid. Anal. Prev.*, vol. 50, pp. 289–297, 2013.

[46] S. H. Khazraee, V. Johnson, and D. Lord, "Bayesian Poisson hierarchical models for crash data analysis: Investigating the impact of model choice on site-specific predictions," *Accid. Anal. Prev.*, vol. 117, pp. 181–195, Aug. 2018, doi: 10.1016/J.AAP.2018.04.016.

[47] D. Lord and F. Mannering, "The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives," *Transp. Res. Part A Policy Pract.*, vol. 44, no. 5, pp.

291–305, 2010, doi: 10.1016/J.TRA.2010.02.001.

[48] F. L. Mannering, V. Shankar, and C. R. Bhat, "Unobserved heterogeneity and the statistical analysis of highway accident data," *Anal. methods Accid. Res.*, vol. 11, pp. 1–16, 2016.

[49] K. El-Basyouny and T. Sayed, "Accident prediction models with random corridor parameters," *Accid. Anal. \& Prev.*, vol. 41, no. 5, pp. 1118–1123, 2009.

[50] J. C. Milton, V. N. Shankar, and F. L. Mannering, "Highway accident severities and the mixed logit model: an exploratory empirical analysis," *Accid. Anal. \& Prev.*, vol. 40, no. 1, pp. 260–266, 2008.

[51] K. S. Conway and T. J. Kniesner, "The important econometric features of a linear regression model with cross-correlated random coefficients," *Econ. Lett.*, vol. 35, no. 2, pp. 143–147, 1991.

[52] N. S. Venkataraman, G. F. Ulfarsson, V. Shankar, J. Oh, and M. Park, "Model of relationship between interstate crash occurrence and geometrics: exploratory insights from random parameter negative binomial approach," *Transp. Res. Rec.*, vol. 2236, no. 1, pp. 41–48, 2011.

[53] E. Coruh, A. Bilgic, and A. Tortum, "Accident analysis with aggregated data: The random parameters negative binomial panel count data model," *Anal. Methods Accid. Res.*, vol. 7, pp. 37–49, 2015.

[54] R. Yu, Y. Xiong, and M. Abdel-Aty, "A correlated random parameter approach to investigate the effects of weather conditions on crash risk for a mountainous freeway," *Transp. Res. Part C Emerg. Technol.*, vol. 50, pp. 68–77, 2015.

[55] T. U. Saeed, T. Hall, H. Baroud, and M. J. Volovski, "Analyzing road crash frequencies with uncorrelated and correlated random-parameters count models: An empirical assessment of multilane highways," *Anal. methods Accid. Res.*, vol. 23, p. 100101, 2019.

[56] D. M. Cerwick, K. Gkritza, M. S. Shaheed, and Z. Hans, "A comparison of the mixed logit and latent class methods for crash severity analysis," *Anal. Methods Accid. Res.*, vol. 3, pp. 11–27, 2014.

[57] Y. Xiong and F. L. Mannering, "The heterogeneous effects of guardian supervision on adolescent driver-injury severities: A finite-mixture random-parameters approach," *Transp. Res. part B Methodol.*, vol. 49, pp. 39–54, 2013.

[58] Y. Xiong, J. L. Tobias, and F. L. Mannering, "The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity," *Transp. Res. part B Methodol.*, vol. 67, pp. 109–128, 2014.

[59] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, "Introducing markov chain monte carlo," *Markov Chain Monte Carlo Pract.*, vol. 1, p. 19, 1996.

[60] H. Goldstein, "Multilevel Statistical Models. Edward Arnold," *London, UK*, 1995.

[61] S.-P. Miaou and D. Lord, "Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods," *Transp. Res. Rec.*, vol. 1840, no. 1, pp. 31–40, 2003.

[62] E. Hauer, "On the estimation of the expected number of accidents," *Accid. Anal. Prev.*, vol. 18, no. 1, pp. 1–12, 1986.

[63] E. Hauer, "Empirical Bayes approach to the estimation of 'unsafety': the multivariate regression method," *Accid. Anal. Prev.*, vol. 24, no. 5, pp. 457–477, 1992.

[64] E. Hauer and B. Persaud, “Common bias in before-and-after accident comparisons and its elimination,” 1983.

[65] B. G. Heydecker and J. Wu, “Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference,” *Adv. Eng. Softw.*, vol. 32, no. 10, pp. 859–869, 2001, doi: https://doi.org/10.1016/S0965-9978(01)00037-0.

[66] Y. C. MacNab, “Bayesian spatial and ecological models for small-area accident and injury analysis,” *Accid. Anal. Prev.*, vol. 36, no. 6, pp. 1019–1028, 2004.

[67] G. Pettet, S. Nannapaneni, B. Stadnick, A. Dubey, and G. Biswas, “Incident analysis and prediction using clustering and Bayesian network,” in *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2017, pp. 1–8, doi: 10.1109/UIC-ATC.2017.8397587.

[68] J. Aguero-Valverde and P. P. Jovanis, “Bayesian Multivariate Poisson Lognormal Models for Crash Severity Modeling and Site Ranking,” *Transp. Res. Rec.*, vol. 2136, no. 1, pp. 82–91, Jan. 2009, doi: 10.3141/2136-10.

[69] S.-P. Miaou and J. J. Song, “Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence,” *Accid. Anal. Prev.*, vol. 37, no. 4, pp. 699–720, 2005.

[70] G. A. Davis and S. Yang, “Bayesian identification of high-risk intersections for older drivers via Gibbs sampling,” *Transp. Res. Rec.*, vol. 1746, no. 1, pp. 84–89, 2001.

[71] P. J. Schlüter, J. J. Deely, and A. J. Nicholson, “Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model,” *J. R. Stat. Soc. Ser. D (The Stat.*, vol. 46, no. 3, pp. 293–316, 1997.

[72] S. Basu and P. Saha, “Regression models of highway traffic crashes: a review of recent research and future research needs,” *Procedia Eng.*, vol. 187, pp. 59–66, 2017.

[73] J. J. Song, M. Ghosh, S. Miaou, and B. Mallick, “Bayesian multivariate spatial models for roadway traffic crash mapping,” *J. Multivar. Anal.*, vol. 97, no. 1, pp. 246–273, 2006, doi: https://doi.org/10.1016/j.jmva.2005.03.007.

[74] B.-J. Park, D. Lord, and J. D. Hart, “Bias properties of Bayesian statistics in finite mixture of negative binomial regression models in crash data analysis,” *Accid. Anal. Prev.*, vol. 42, no. 2, pp. 741–749, 2010.

[75] M. Abdel-Aty, A. Pande, A. Das, and W. J. Knibbe, “Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems,” *Transp. Res. Rec.*, vol. 2083, no. 1, pp. 153–161, Jan. 2008, doi: 10.3141/2083-18.

[76] R. Yu and M. Abdel-Aty, “Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data,” *Saf. Sci.*, vol. 63, pp. 50–56, 2014, doi: 10.1016/j.ssci.2013.10.012.

[77] Y. Xie, D. Lord, and Y. Zhang, “Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis,” *Accid. Anal. Prev.*, vol. 39, no. 5, pp. 922–933, Sep. 2007, doi: 10.1016/J.AAP.2006.12.014.

[78] S. Basak, A. Dubey, and L. Bruno, “Analyzing the Cascading Effect of Traffic Congestion Using LSTM Networks,” *Proc. - 2019 IEEE Int. Conf. Big Data, Big Data 2019*, pp. 2144–2153, 2019, doi: 10.1109/BigData47090.2019.9005995.

[79] Z. Iqbal, M. I. Khan, S. Hussain, and A. Habib, “An Efficient Traffic Incident Detection and Classification Framework by Leveraging the Efficacy of Model Stacking,” *Complexity*, vol. 2021, 2021.

[80] J. Tang, J. Liang, C. Han, Z. Li, and H. Huang, “Crash injury severity analysis using a two-layer Stacking framework,” *Accid. Anal. \& Prev.*, vol. 122, pp. 226–238, 2019.

[81] J. Xiao, “SVM and KNN ensemble learning for traffic incident detection,” *Phys. A Stat. Mech. its Appl.*, vol. 517, pp. 29–35, 2019.

[82] Z. Ma, G. Mei, and S. Cuomo, “An analytic framework using deep learning for prediction of traffic accident injury severity based on contributing factors,” *Accid. Anal. \& Prev.*, vol. 160, p. 106322, 2021.

[83] A. Behura and A. Behura, “Road Accident Prediction and Feature Analysis by Using Deep Learning,” *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, 2020, doi: 10.1109/ICCCNT49239.2020.9225336.

[84] D. Singh and C. K. Mohan, “Deep spatio-temporal representation for detection of road accidents using stacked autoencoder,” *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 879–887, 2018.

[85] T. Chen, X. Shi, and Y. D. Wong, “Key feature selection and risk prediction for lane-changing behaviors based on vehicles’ trajectory data,” *Accid. Anal. \& Prev.*, vol. 129, pp. 156–169, 2019.

[86] Y. Qi, B. L. Smith, and J. Guo, “Freeway accident likelihood prediction using a panel data analysis approach,” *J. Transp. Eng.*, vol. 133, no. 3, pp. 149–156, 2007.

[87] “INRIX.” https://inrix.com/.

[88] J. MacQueen and others, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1967, vol. 1, no. 14, pp. 281–297.

[89] C.-S. Li, J.-C. Lu, J. Park, K. Kim, P. A. Brinkley, and J. P. Peterson, “Multivariate zero-inflated Poisson models and their applications,” *Technometrics*, vol. 41, no. 1, pp. 29–38, 1999.

[90] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[91] C. Chen, A. Liaw, and L. Breiman, “Using Random Forest to Learn Imbalanced Data,” *Univ. California, Berkeley*, 2004.

[92] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[93] A. B. Calvo and D. H. Marks, “Location of health care facilities: an analytical approach,” *Socioecon. Plann. Sci.*, vol. 7, no. 5, pp. 407–422, 1973.

[94] D. Serra and V. Marianov, “The p-median problem in a changing network: the case of Barcelona,” *Locat. Sci.*, vol. 6, no. 1–4, pp. 383–394, 1998.

[95] M. Dzator and J. Dzator, “An effective heuristic for the p-median problem with application to ambulance location,” *Opsearch*, vol. 50, no. 1, pp. 60–74, 2013.

[96] L. Caccetta, M. Dzator, and others, “Heuristic methods for locating emergency facilities,” 2005.

[97] O. Kariv and S. L. Hakimi, “An algorithmic approach to network location problems. I: The p-centers,” *SIAM J. Appl. Math.*, vol. 37, no. 3, pp. 513–538, 1979.

[98] M. S. Daskin, *Network and discrete location: models, algorithms, and applications*. John Wiley \& Sons, 1995.

[99] A. F. Agarap, “Deep learning using rectified linear units (relu),” *arXiv Prepr. arXiv1803.08375*, 2018.

[100] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv Prepr. arXiv1412.6980*, 2014.

[101] R. Sánchez-Mangas, A. Garc\'\ia-Ferrrer, A. De Juan, and A. M. Arroyo, “The probability of death in road traffic accidents. How important is a quick medical response?,” *Accid. Anal. \& Prev.*, vol. 42, no. 4, pp. 1048–1056, 2010.

Appendices

1) Docker

Docker is a platform for OS-level virtualization to deliver our pipeline. The pipeline and all required libraries are provided in an image file. Then docker uses that image to create a container (a virtual OS-level environment) and run the intended commands. In this subsection we briefly review how to run a docker and what each component in the commands provided before means.

First, we use the following command in the terminal to create an image called prediction_engine_1.

```
Docker build -t prediction_engine_1.
```

Then there are four functions available inside of the image

run_dataprep.py: it runs the *Data Preparation* module.

run_training.py: it runs the *Training Models* module.

run_prediciton.py: it runs the *Prediction* module.

run_simulation.py: it runs the *Simulation* module.

For each command, three items should be defined, i.e., config file using the flag c or config, the location of input files using the flag i or input, and the location of the output files (the location you want the module to save the results or outputs,) using the flag o or output. The locations are the location of the files inside of the container. For example, they can be defined as follows:

```
-c /app/etc/config.conf

-i /app/data

-o /app/output1
```

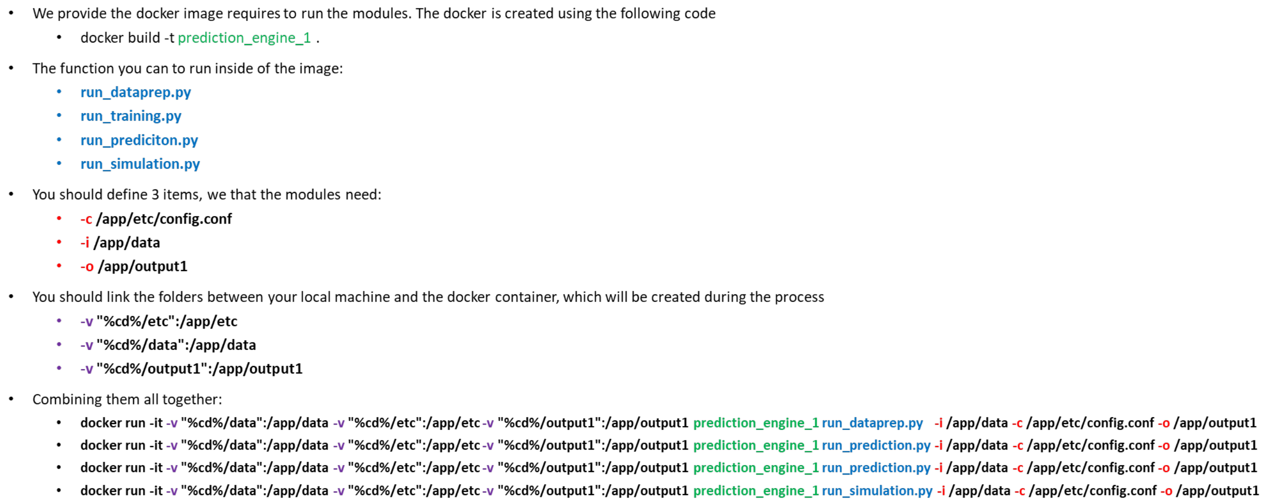
Basically, the config file and input files should be fed to the model. Therefore, the image file does not include them. In general, they are provided in separate files and mounted (link the folders between your local machine and the docker container) by the user. For mounting, two parameters should be defined. The location of the file in the location machine and the location of the file in container. By doing so, we link these folders and any changes in one of them is reflected in the other one as well. The same process should be done for the output folder to extract the output files and results created by the container. Consequently, the three sample examples for each are as follows:

```
-v "%cd%/etc":/app/etc

-v "%cd%/data":/app/data

-v "%cd%/output1":/app/output1
```

Figure A.1 summarizes the aforementioned commands.



- We provide the docker image requires to run the modules. The docker is created using the following code
  - docker build -t prediction_engine_1 .
- The function you can to run inside of the image:
  - run_dataprep.py
  - run_training.py
  - run_prediciton.py
  - run_simulation.py
- You should define 3 items, we that the modules need:
  - -c /app/etc/config.conf
  - -i /app/data
  - -o /app/output1
- You should link the folders between your local machine and the docker container, which will be created during the process
  - -v "%cd%/etc":/app/etc
  - -v "%cd%/data":/app/data
  - -v "%cd%/output1":/app/output1
- Combining them all together:
  - docker run -it -v "%cd%/data":/app/data -v "%cd%/etc":/app/etc -v "%cd%/output1":/app/output1 prediction_engine_1 run_dataprep.py -i /app/data -c /app/etc/config.conf -o /app/output1
  - docker run -it -v "%cd%/data":/app/data -v "%cd%/etc":/app/etc -v "%cd%/output1":/app/output1 prediction_engine_1 run_prediction.py -i /app/data -c /app/etc/config.conf -o /app/output1
  - docker run -it -v "%cd%/data":/app/data -v "%cd%/etc":/app/etc -v "%cd%/output1":/app/output1 prediction_engine_1 run_prediction.py -i /app/data -c /app/etc/config.conf -o /app/output1
  - docker run -it -v "%cd%/data":/app/data -v "%cd%/etc":/app/etc -v "%cd%/output1":/app/output1 prediction_engine_1 run_simulation.py -i /app/data -c /app/etc/config.conf -o /app/output1

Figure A.1 Summary of the docker codes

2) Data Cleaning

This section is dedicated to briefly mention some of the details of the challenges we have faced regarding collecting, cleaning, combining and aggregating different datasets required for incident prediction.

The first step is understanding and cleaning the incident dataset available through TDOT. Here, two issues regarding incident dataset are reviewed. While, the quality of the data has improved during the time, some of the main sources of erroneous sample data are worth to be mentioned. Figure A.2 shows the spatial distribution of incidents in the raw dataset according to the latitude and longitude of the samples. There are plenty of sample outside of the boundary of Tennessee while the samples are supposed to be limited to traffic incidents occurred in the state Tennessee. The temporal information of the samples is also erroneous. While the original format of the feature regarding the time of accident is 24-hour, it suddenly changes to 12-hour format without keeping the am/pm label. This error reduces the rate of accidents for any time frame after 12:59 pm to zero.

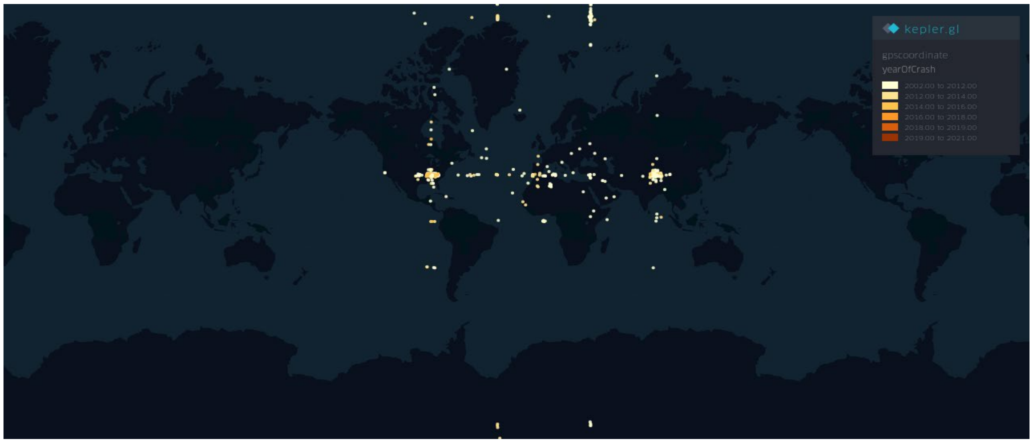


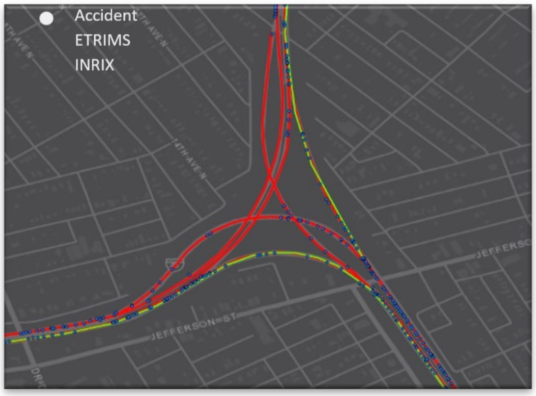
Figure A.2 Spatial distribution of incidents in the raw dataset

Then, we should choose a geographical framework for the highway segmentation. This defines the location and geometry of each roadway segment in our problem. There are two main alternatives: ETRIMS (Enhanced Tennessee Roadway Information Management System) and INRIX. Incident dataset includes a feature that defines the relevant highway segment based on ETRIMS to each segment. This obviates the need for mapping incidents to segments. While ETRIMS can be useful for low spatial resolution incident prediction, it has a noticeable drawback, highlighted by Figure A.3. ETRIMS represents two opposite direction of a highway as well as all of the ramps and connectors between them by one segment. Obviously, this approach can be problem when accidents should be studied in high spatial resolution. Furthermore, this drawback limits us in using traffic flow information since ETRIMS combines all segments in one. Therefore, we choose INRIX for the highway segmentation framework. INRIX can also provide the traffic data.

One of the features required to be investigated is elevation and the change in the elevation of the roadway segments. Google Cloud Elevation API[6] and United States Geological Survey (USGS) Elevation API[7] are the two main available APIs for extracting the altitute/elevation of the a point using its latitute and longitute. However, they work as a one to one mapping function, which means for each pair and of latitute and longitute they return one value for the elevation. However, in the case of a brige, a point on the deck of the bridge and a point on the road passing under the bridge have two different elevation while they have the same coodinates, as shown in Figure A.4. We address this problem by limiting the maximum allowable slope to 5%. The points violating this limit are detected as an anomoloy and replaced by interploated values.

We choose Weatherbit over Darsky due to lower percentage of missing values. However, the collected weather dataset still includes some missing values, in particular in some imporant weather features such as temperatures. It can happen due to various reasons.

Figure A.3 ETRIMS (thick green) and INRIX (thin red) highway segments and location of accidents



[6] https://developers.google.com/maps/documentation/elevation/start

[7] https://nationalmap.gov/epqs/

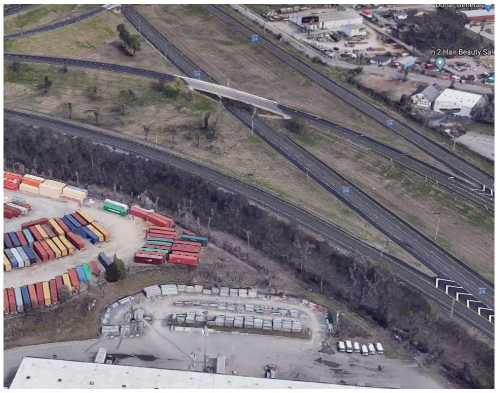


Figure A.4 Caveat of using Google Maps Elevation API for extracting the elevation a) 3D perspective of an unleveled intersection b) 3D overview of the same intersection using the elevation extracted from Google Maps Elevation API

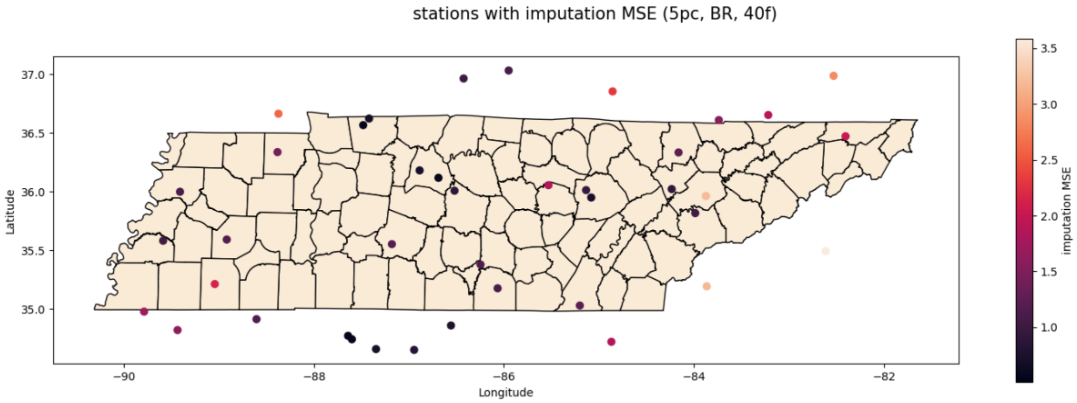


Figure A.5 Distribution of station-wise imputation MSEs for temperature plotted on the county-level map of Tennessee using the geographic location of each station (using Bayesianridge with 40 features and 5% NA values)

3) Feature Analysis

This section summarizes figures generated by the method discussed in Feature Analysis section.

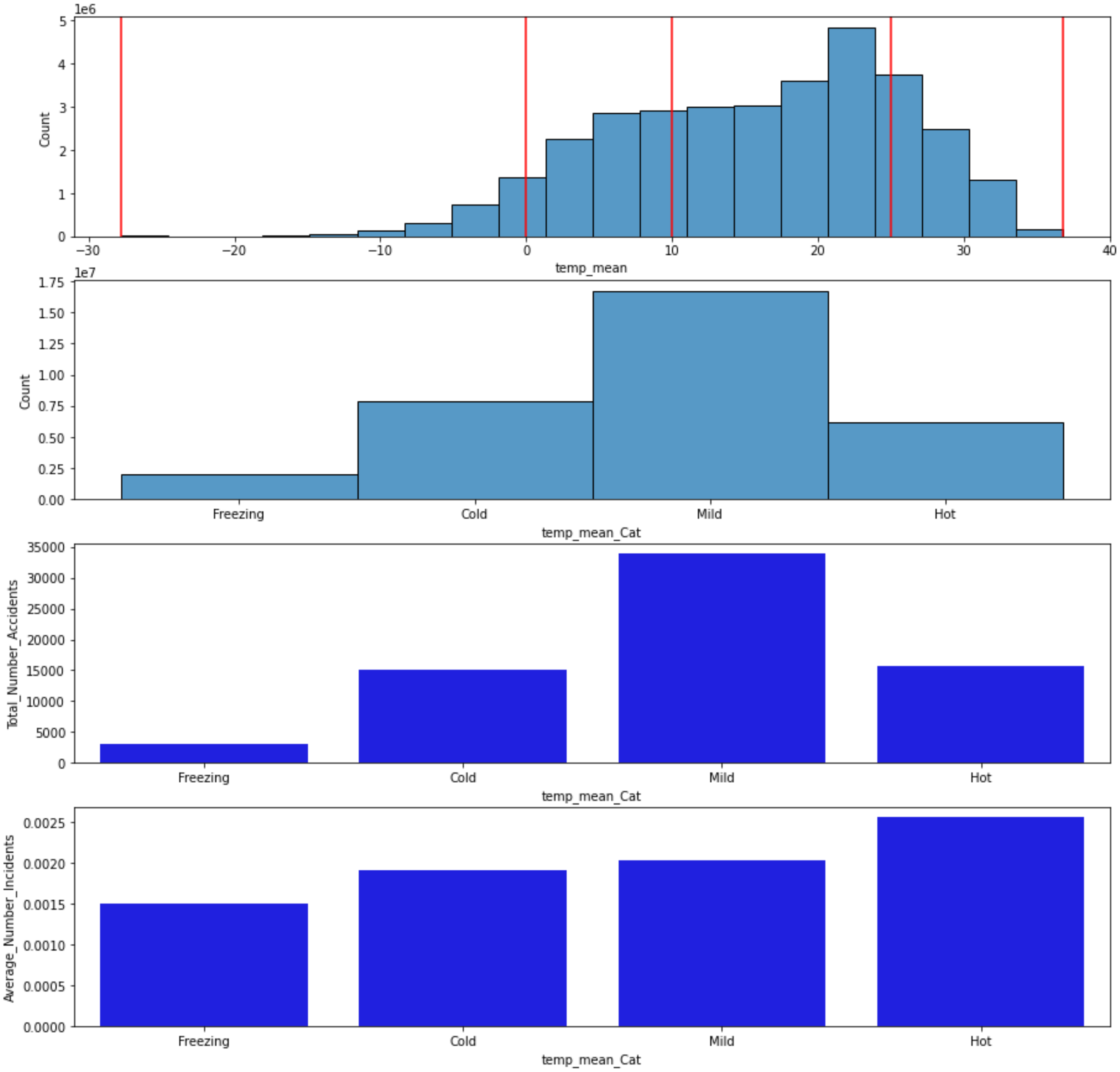


Figure A.6 Feature analysis; estimation of importance of temperature in accident occurrence

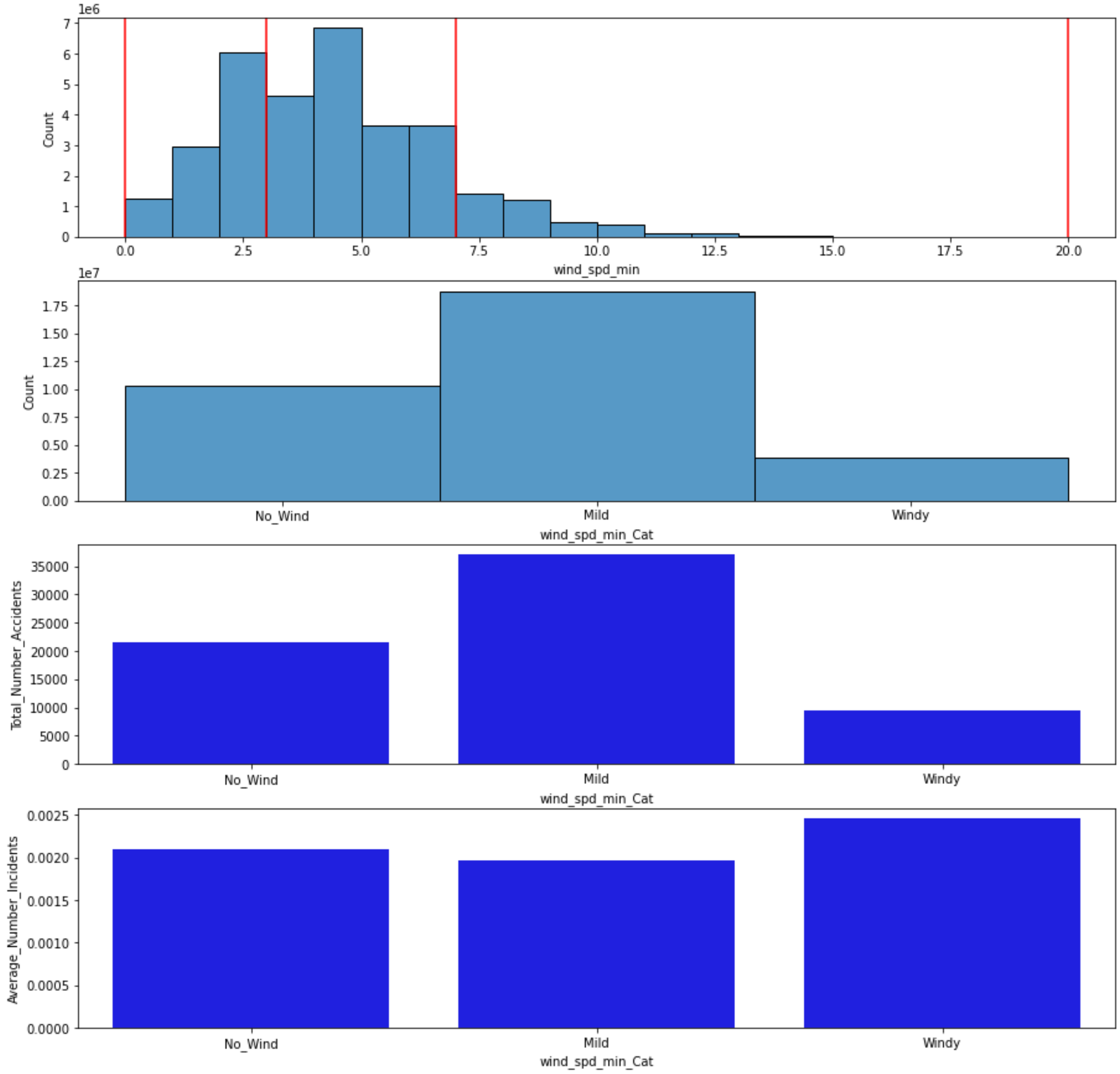


Figure A.7 Feature analysis; estimation of importance of wind in accident occurrence

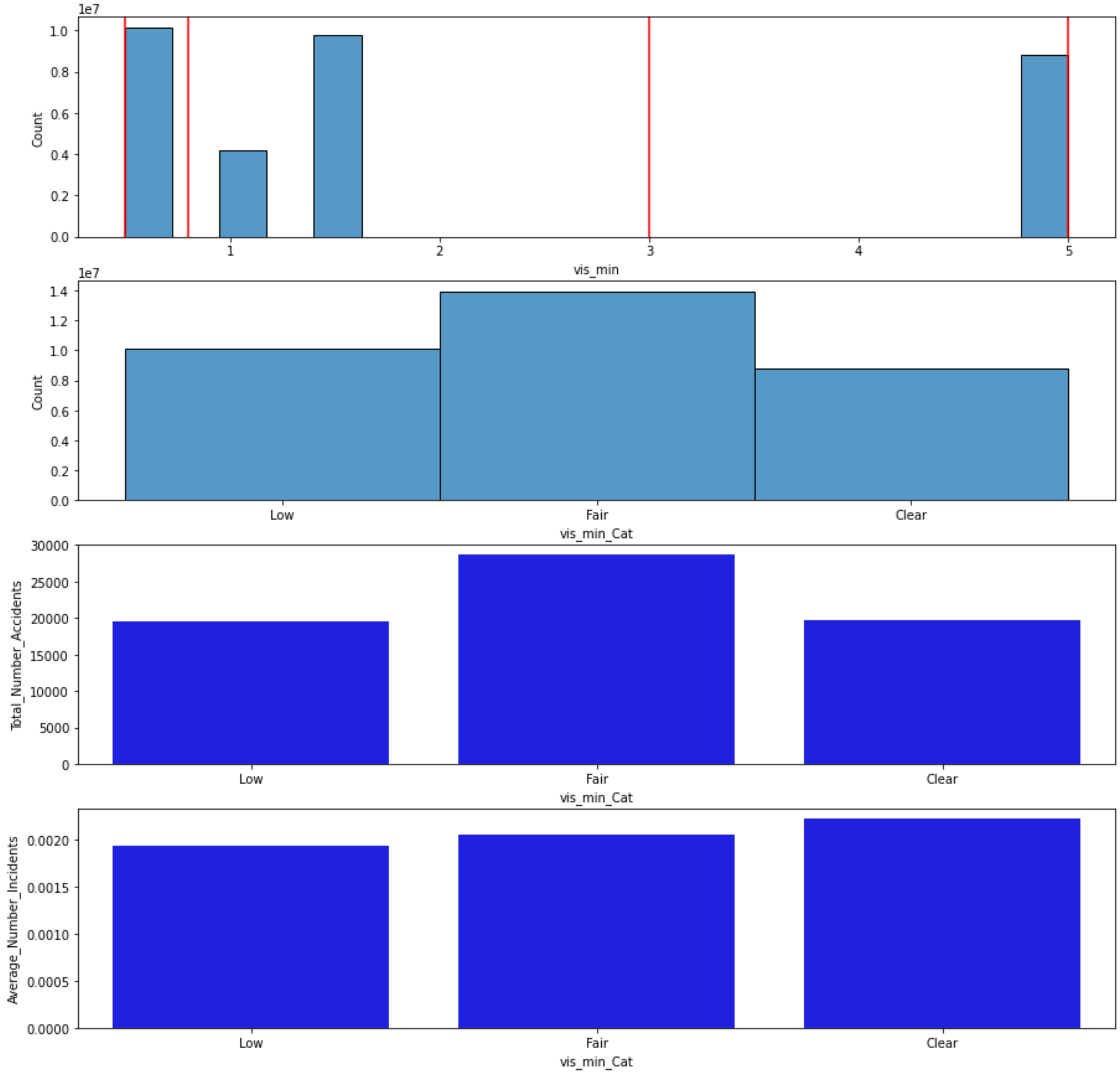


Figure A.8 Feature analysis; estimation of importance of visibility in accident occurrence

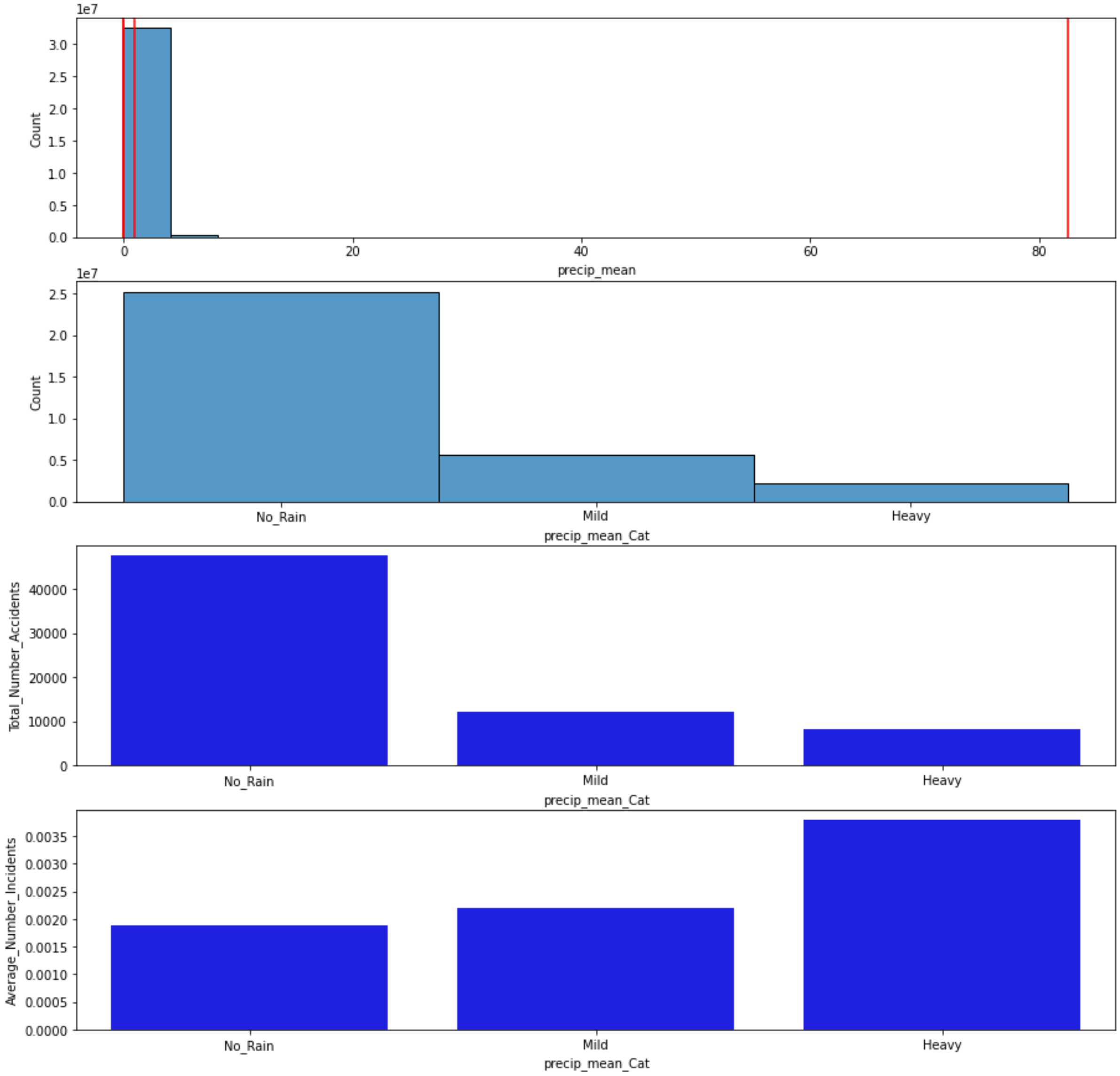


Figure A.9 Feature analysis; estimation of importance of precipitation in accident occurrence

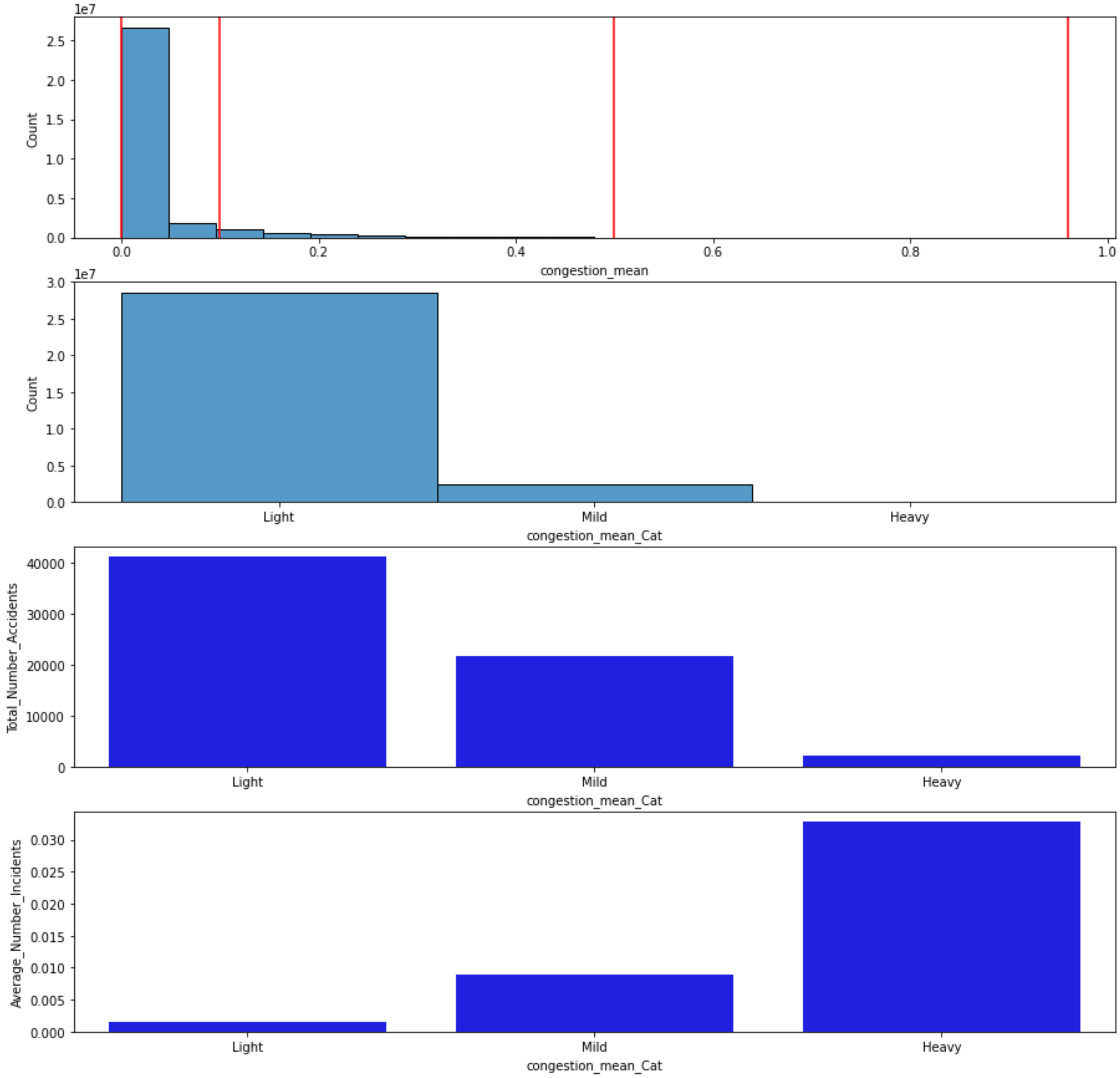


Figure A.10 Feature analysis; estimation of importance of congestion in accident occurrence

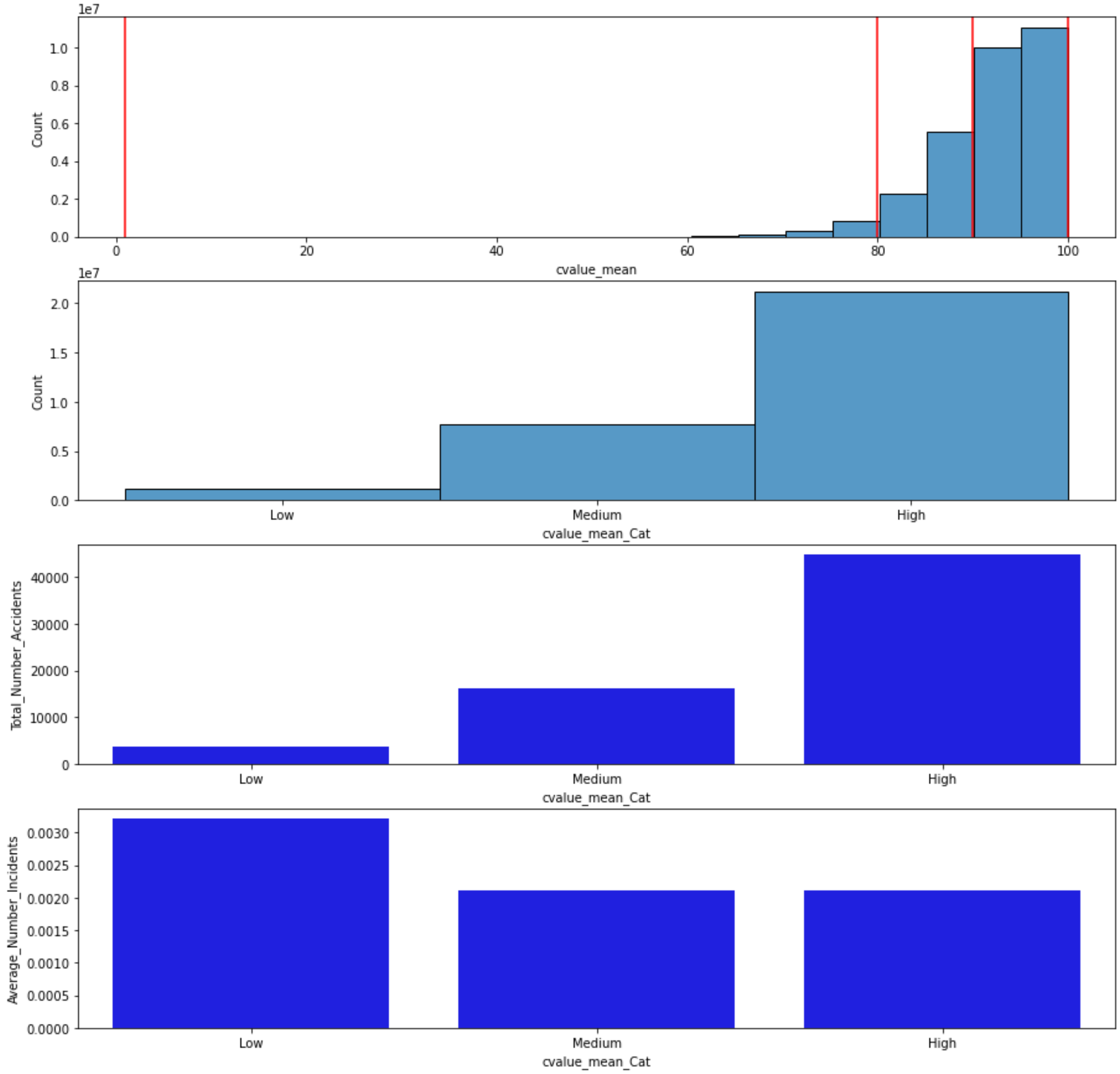


Figure A.11 Feature analysis; estimation of importance of C-value in accident occurrence

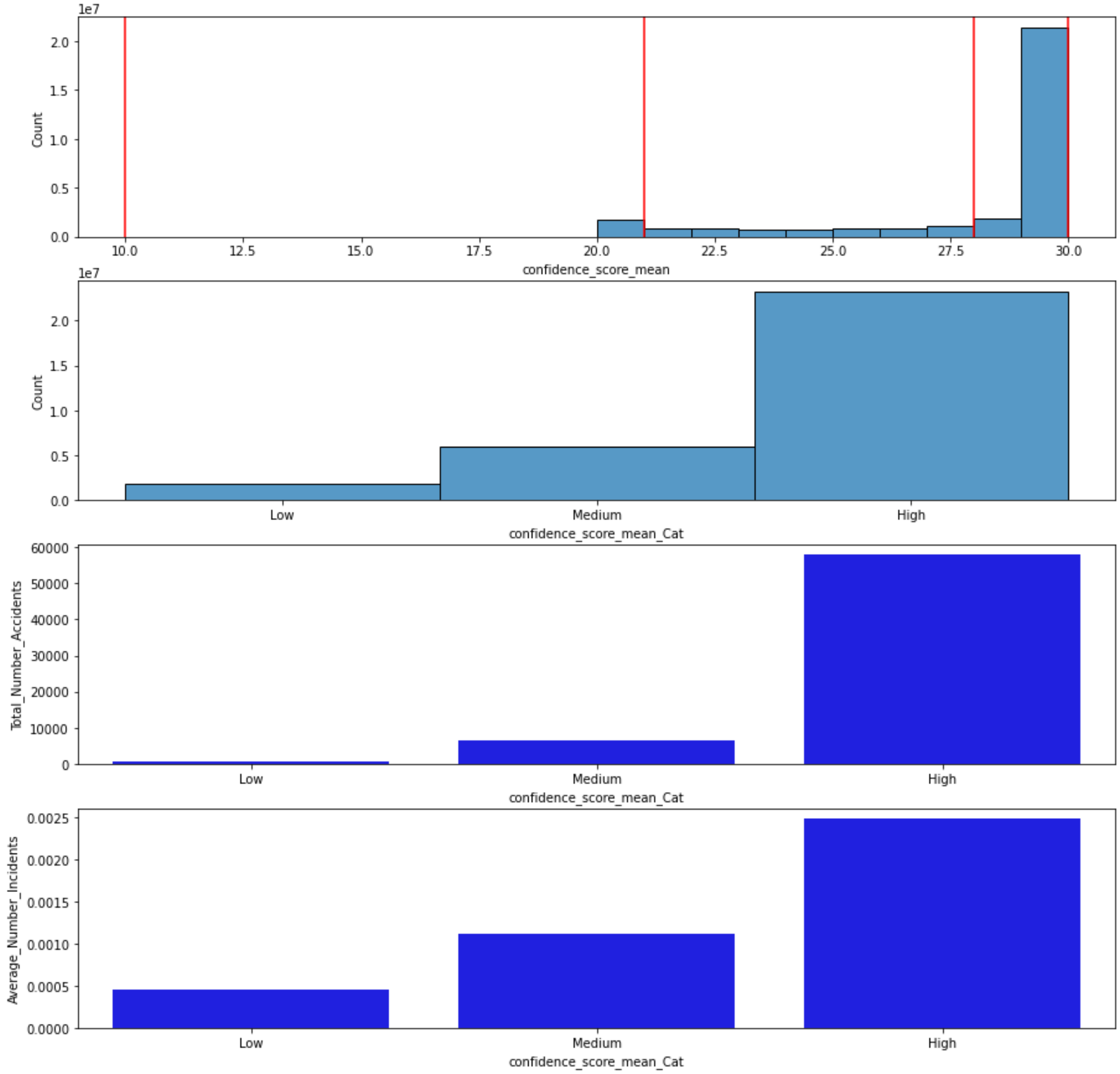


Figure A.12 Feature analysis; estimation of importance of confidence score in accident occurrence

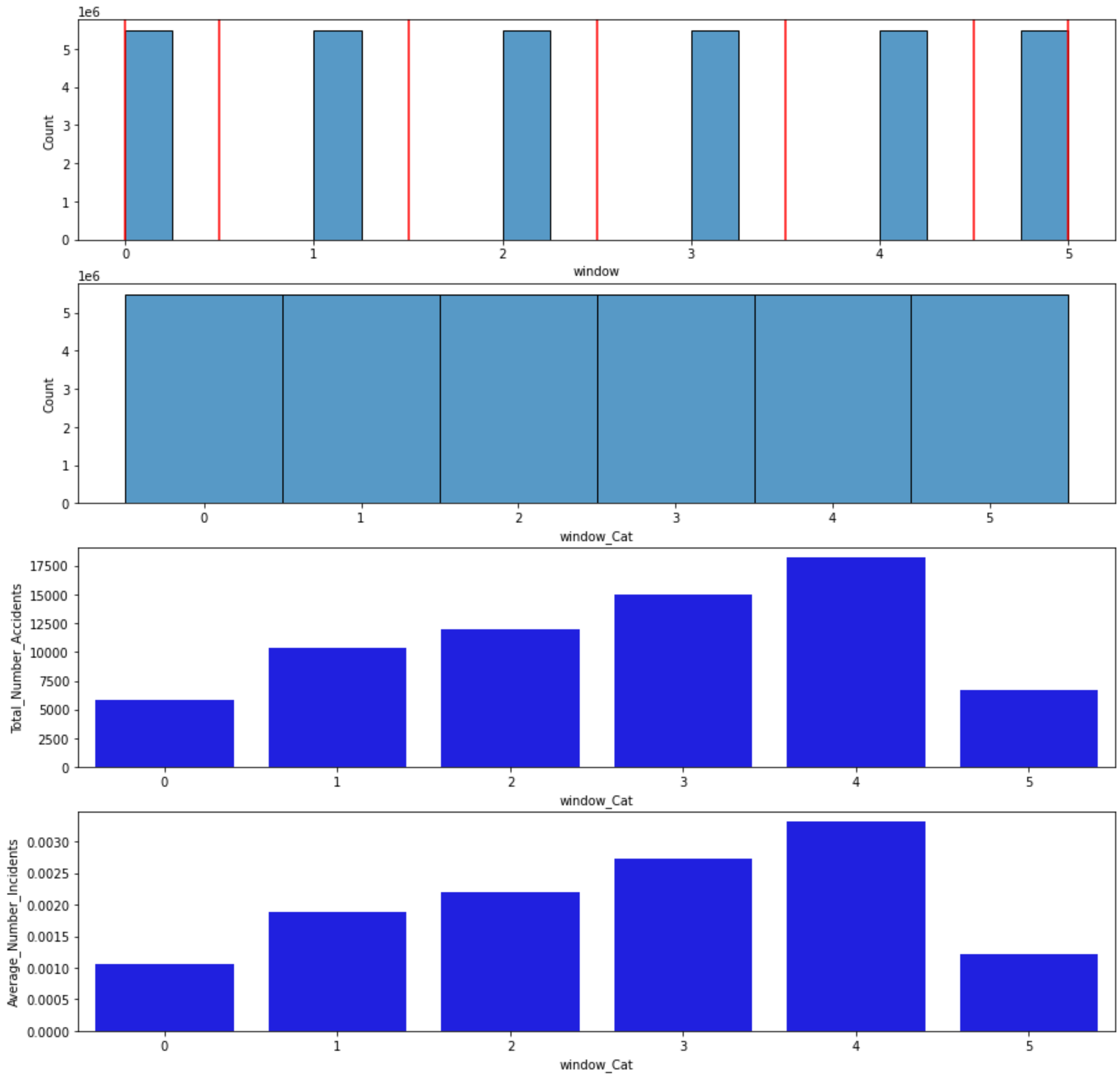


Figure A.13 Feature analysis; estimation of importance of time of the day (window) in accident occurrence

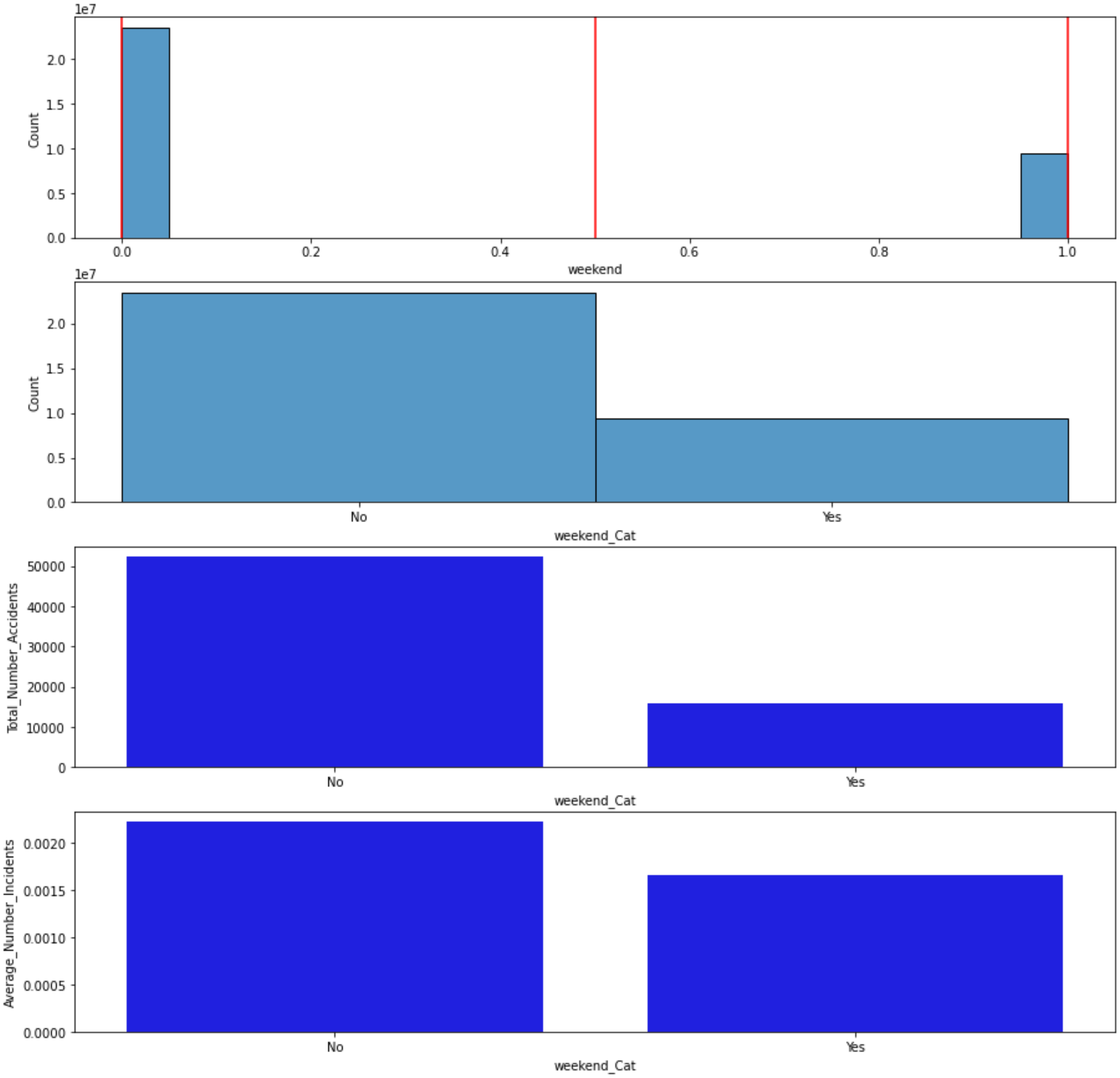


Figure A.14 Feature analysis; estimation of importance of weekend in accident occurrence

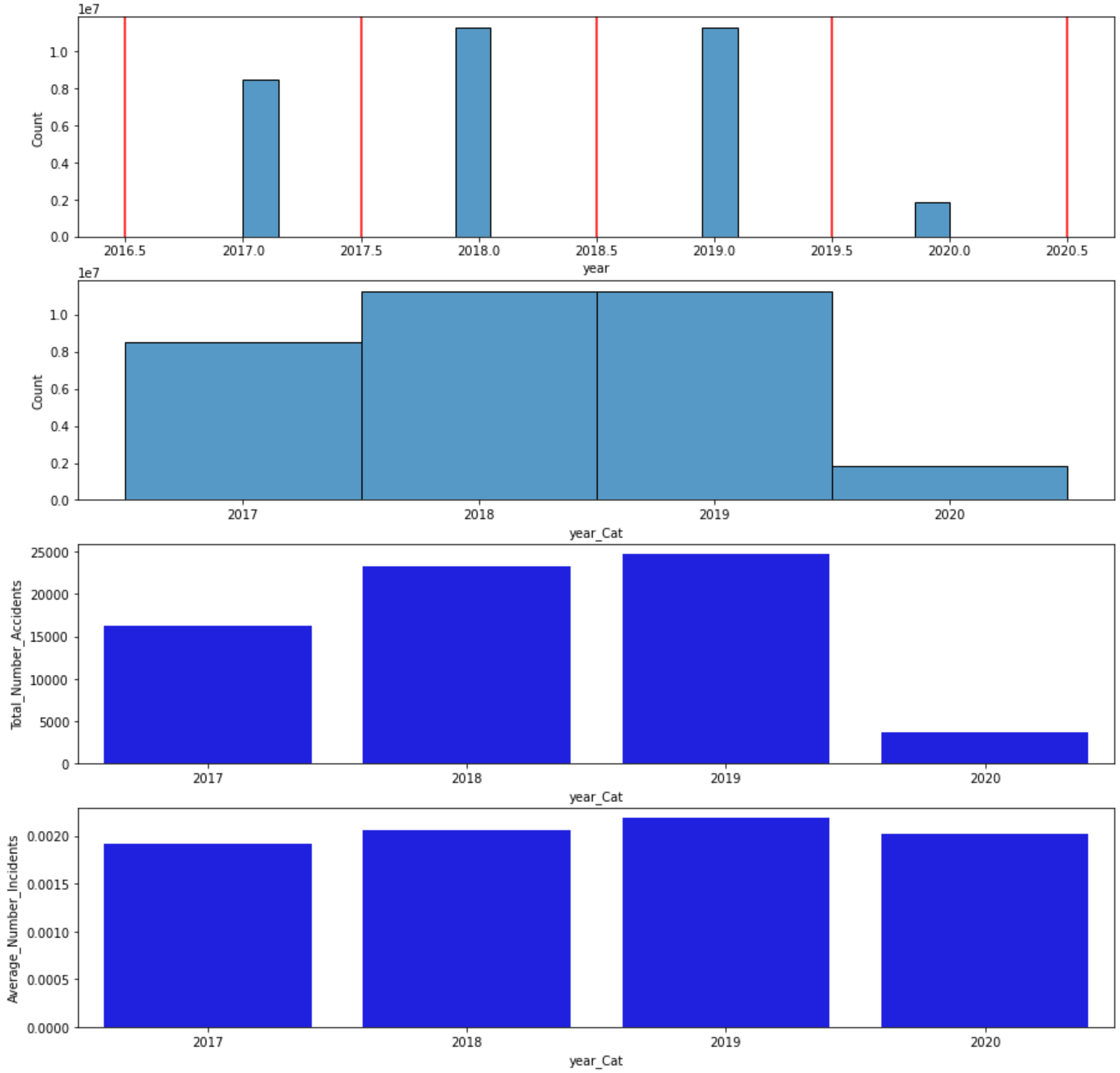


Figure A.15 Feature analysis; estimation of importance of year in accident occurrence

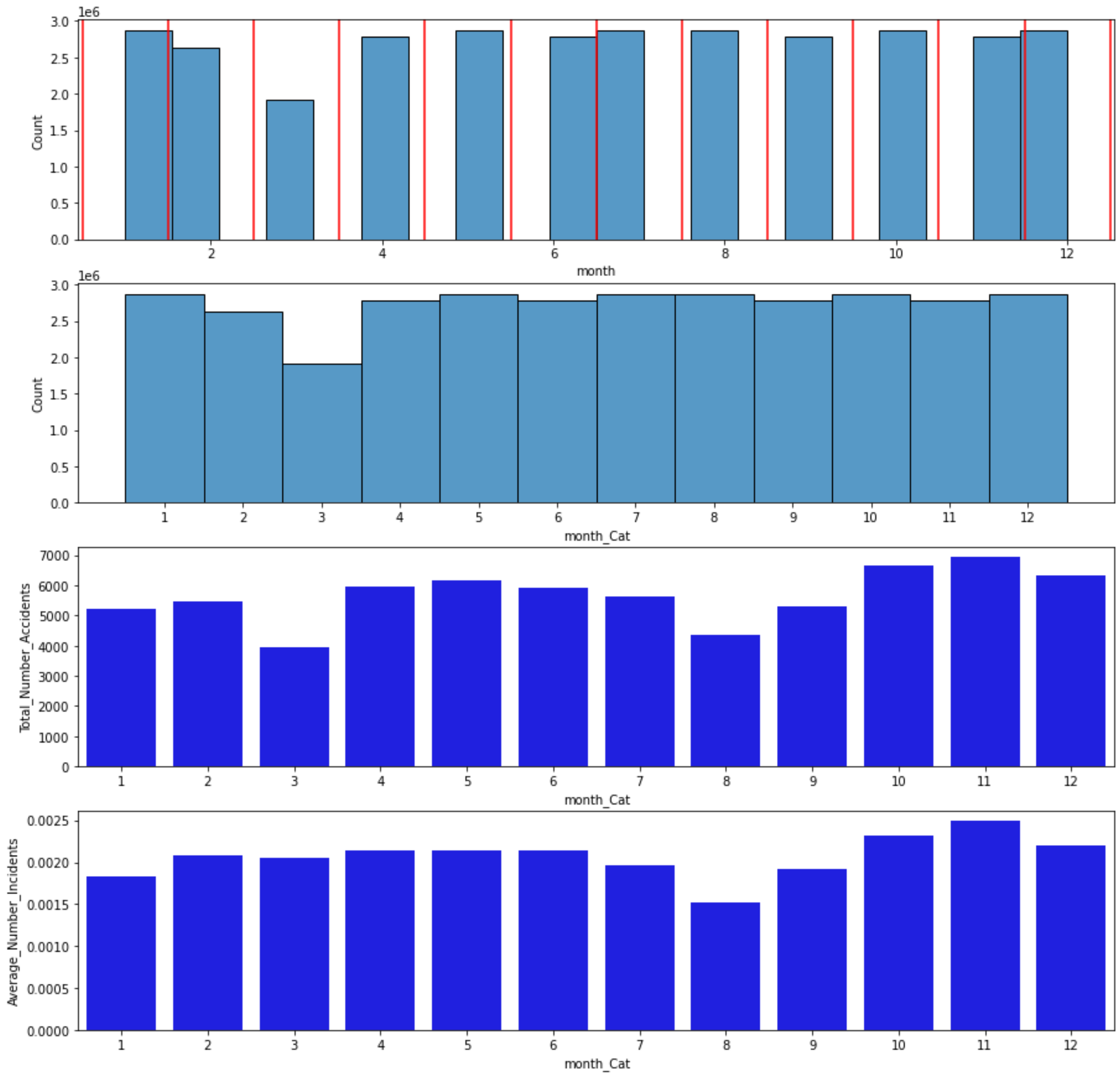


Figure A.16 Feature analysis; estimation of importance of month in accident occurrence

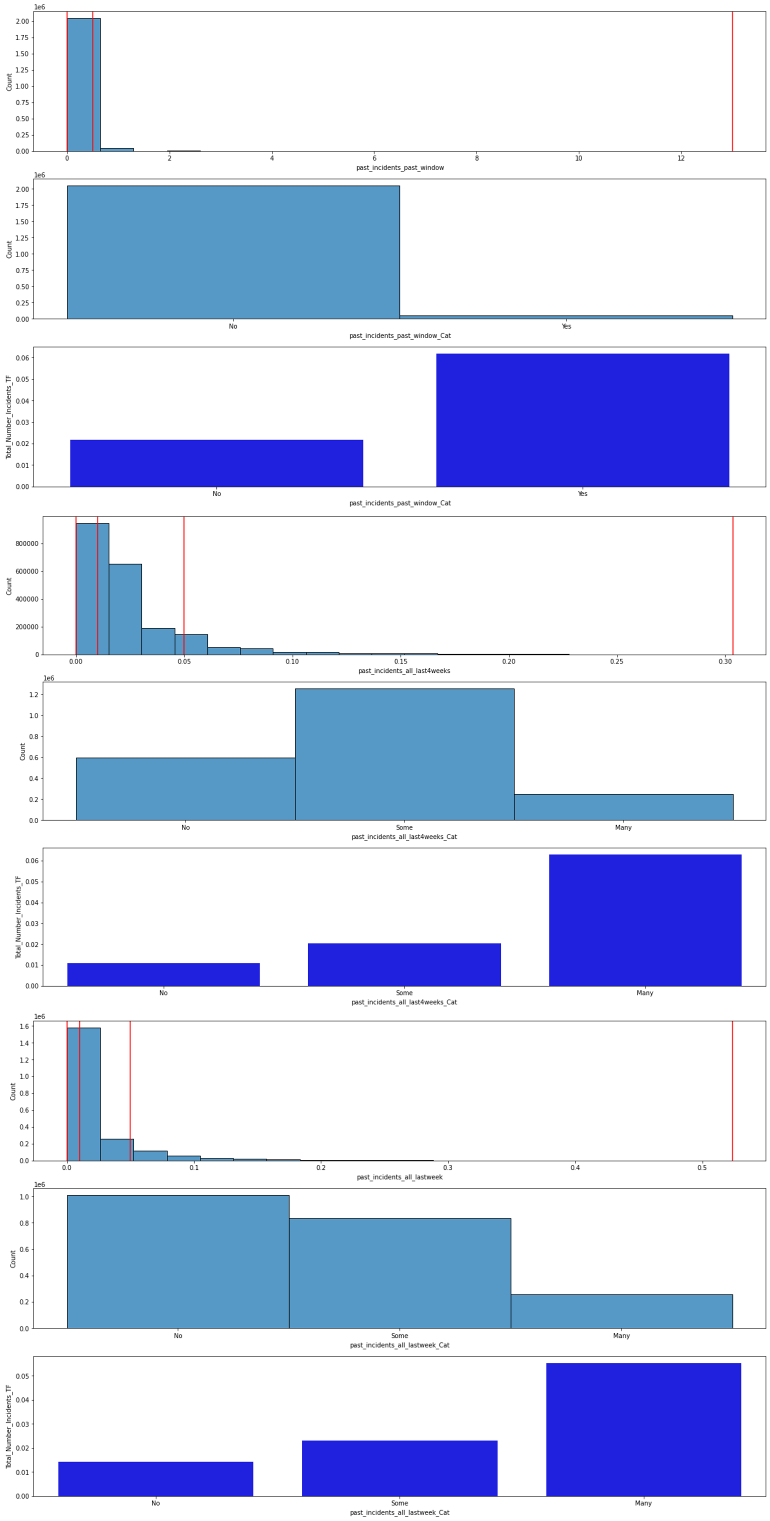


Figure A.17 Feature analysis; estimation of importance of historical incidents in accident occurrence