Building an Accident Risk Prediction Model Using Machine Learning Algorithms on Heterogeneous Sparse Data

Anirudh Srinivasan
Arizona State University,
Tempe, USA
anirudhs@asu.edu

Manohar Akula
Arizona State University,
Tempe, USA
makula@asu.edu

Pranav Murali
Arizona State University,
Tempe, USA
pmural15@asu.edu

Srikanth Sampath
Arizona State University,
Tempe, USA
ssampa28@asu.edu

ABSTRACT

Minimizing road accidents is a paramount challenge that is faced by the United States today. The CDC claims that road accidents is the leading cause of deaths in the country [1]. The country witnesses on average an alarming 100 deaths each day due to this [2]. Therefore it is of utmost importance to study and analyze road traffic accidents to appropriately handle and thus reduce the number of accidents. Assessing the severity of accidents is an important aspect that must be examined in detail to address this issue. Existing studies to analyze and predict the risk of accidents face the challenge of either having to use small scale datasets with limited coverage or used large scale datasets which weren't updated frequently or did not have important contextual information to fully assess the situation during the accident. In this project, we utilize a dataset [3] that has been created to capture accident data on a country wide level including important aspects like weather, points-of-interest and severity of accident ,.etc. Using this dataset, we implement state-of-the-art machine learning (ML) techniques to predict the severity of an accident. Our solution builds on a novel and thorough dataset that has recorded information about all accidents on a country wide scale which currently has around 3 million entries. Through this prediction, we aim to analyze the conditions which are likely to cause fatal accidents and plan accordingly and reduce such incidents in future. Also using these predictions, we aim to better equip emergency response teams in locations based on severity which would aid the teams to respond quickly and also be better prepared to handle similar situations in the future. This project shows that such huge datasets could be leveraged to predict the 'Severity' accurately using the models used.

KEY WORDS: Risk Prediction, Neural Structured Learning, Hummingbird, Adversarial Regularization

INTRODUCTION

Along with the advent of modern technology and road transport across the country becoming essential for personal and commercial purposes, we see an alarming increase in the number of road accidents day by day.

The US witnessed 38,680 deaths in the year 2020. This represents an increase of 7.2% from 2019 [3]. This was despite the fact that people travelled and drove far lesser compared to previous years due to the global pandemic. This statistic alone signifies how much accident related deaths and injuries have spread like cancer in our lives causing life-long trauma for the people involved.

In the past, although there have been many studies conducted and researches carried out to analyze the cause and effects of accidents, there were huge data constraints faced by them. These studies were mainly limited to using datasets that had information with very limited coverage. Findings from these datasets could not be extrapolated to large-scale level as they would not be accurate or relevant. On the contrary, any large scale dataset that was previously used in studies did not contain important, relevant information like weather, points-of-interest, severity of the accident which would be necessary to map out relations and trends for predicting road accidents and plans to reduce them. With the increased usage of technology in everyday life, it is therefore of utmost importance to accurately capture all possible relevant data with regards to road accidents to analyze them and accurately predict such incidents and be prepared with the necessary course of action to avoid fatalities.

Some of the limitations of the past datasets included limited coverage (just one city, small number of road segments)[4, 5, 6]; datasets that were dependent on a wide range of data attributes that may not be constantly available for all regions [7,8,9]. Even large scale datasets could not be obtained for real time predictions leading to obstacles in model building and hence produce subpar predictions with a lot of uncertainties. These were the main shortcomings of previous studies.

To address these challenges we propose a solution to build a robust model that would be able to accurately predict the *Severity* of an accident using data that has information about a wide range of data events like traffic events which include traffic congestion and hazards; weather events like temperature and wind speed; specific points-of-interest like traffic signals and junction; time information like week, hour and period of the day. Using such information in ML models would require the model to handle such large volumes data that would keep increasing as time progresses. Building a model that is capable of handling such volumes of data will enable the predictions to be more accurate and relevant to the current situation. The problem at hand demands such kinds of accuracies as it will directly impact the lives of people across the country.

We are going to be using a state-of-the-art ML model named "Neural Structured Learning" (NSL) for this project. It has improved upon the previous neural network training models using structured data, particularly when the amount of labeled data is relatively small. This technique also helps in building a robust model against any misleading predictions.

Next, we are also going to be making use of another novel ML tool called Hummingbird. Training of models can be vastly enhanced by integrating hummingbird with the existing models through tensor computations. Using this, we can enhance the computation power and ability to process larger datasets with millions of records much faster without having to re-engineer them.

METHODOLOGY

In this project, we build robust models that would be able to handle large amounts of data that would be generated year-on-year. To implement such a model we use the Neural Structured Learning model and the Hummingbird toolset. We then compare the results of these models against various traditional supervised models that have been mentioned below. This section explains in detail the working of the models, techniques and tools that have been implemented.

A. Baseline Models

This section outlines the various baselines models that have been implemented in our project. We then compare the accuracies of these models against the state-of-the-art models that we have outlined and implemented later

Logistic Regression: This model is used for classification problems and to determine if a new data sample would correctly suit one of the the categorical labels.In this project, we use logistic regression on the accident dataset and implement it as one of the baseline models to perform classification.

K-Nearest Neighbor: It is a supervised ML model and used for regression and classification. It can be used for instance based learning where training data is used to predict output for future instances. On the accident datasets, this model classifies the severity of the accident based on the data fed to the model.

Decision Tree: A supervised ML algorithm that learns from training data to predict the class or label of the target variable using simple decision rules. In this project, we predict the target variable, *'Accident Severity'* using this model.

Random Forest: This is a ML model which is an extension of the decision tree model where the output is predicted based on the inputs from different decision trees.

MLP Classifier: This model relies on neural networks to perform classification tasks and is a type of feed forward neural network.

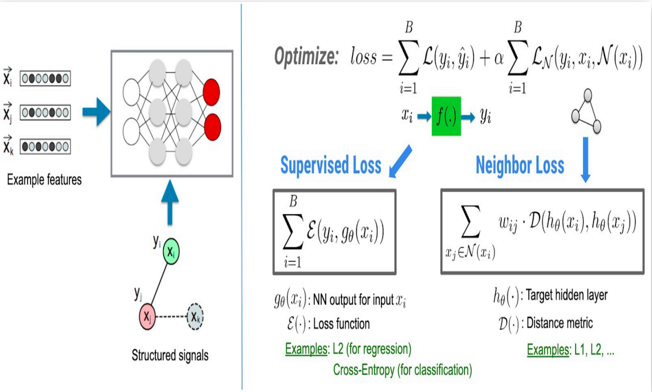
B. Neural Structured Learning Model

Introduced by Google in September 2019, Neural Structured Learning (NSL) is a model that has been developed to leverage the structured features in the training data for deep neural networks. This technique has a wide range of applications and can be implemented for different types of neural networks like deep, convolutional or feed-forward neural networks.

These structured signals are used to regularize the training of a neural network, forcing the model to learn accurate predictions by minimizing the supervised loss while at the same time maintaining the input similarity by minimizing the neighbor loss [10].

Initially the training samples are reinforced to include structured signals which can be explicitly provided or can be induced during adversarial learning. The reinforced training samples which include both the original samples and their corresponding neighbors are fed to the neural network. Along with the final loss, the neighbor loss is also calculated and we arrive at a cumulative weighted loss.

Higher accuracy is a major advantage that the NSL model provides. This is because the structured signals as part of the sampled data set can provide more information than feature inputs. This model is also robust in nature as it is able to withstand adversarial perturbations. Another significant aspect of this model is that it allows the network to train using the labeled data as in the traditional supervised learning models and at the same time drives the network to learn similar hidden representation for the 'neighboring sample' that may or may not have labels[11].



$\vec{x}_i$
$\vec{x}_j$
$\vec{x}_k$
Example features
y_i
y_j
x_i
x_j
Structured signals
Optimize: $loss = \sum_{i=1}^{B} \mathcal{L}(y_i, \hat{y}_i) + \alpha \sum_{i=1}^{B} \mathcal{L}_{\mathcal{N}}(y_i, x_i, \mathcal{N}(x_i))$
$x_i \rightarrow f(\cdot) \rightarrow y_i$
Supervised Loss
Neighbor Loss
$\sum_{i=1}^{B} \mathcal{E}(y_i, g_\theta(x_i))$
$\sum_{x_j \in \mathcal{N}(x_i)} w_{ij} \cdot \mathcal{D}(h_\theta(x_i), h_\theta(x_j))$
$g_\theta(x_i)$: NN output for input x_i
$\mathcal{E}(\cdot)$: Loss function
Examples: L2 (for regression)
Cross-Entropy (for classification)
$h_\theta(\cdot)$: Target hidden layer
$\mathcal{D}(\cdot)$: Distance metric
Examples: L1, L2, ...

Figure 1:With structured signals as input, the loss function is a sum of supervised and neighbor loss[11]

C. Hummingbird

Introduced by Microsoft in 2020, it is a novel ML tool that would help researchers and developers a great deal in dealing with bottlenecks that arise due to computational capabilities of ML models. It helps in the integration of models to the PyTorch framework that can aid in running the models on GPUs and obtain

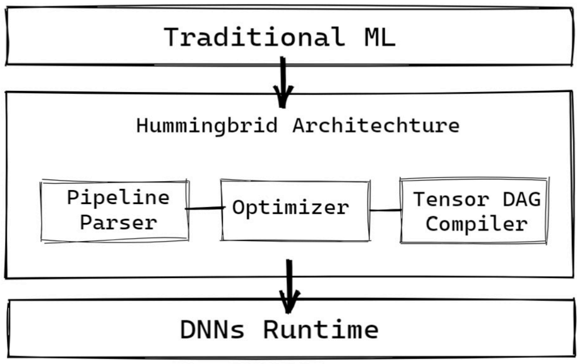


Fig 2: High Level Architecture of Hummingbird [12]

better performances. Hummingbird helps in compiling ML pipelines in the form of tensor computations that help deal with neural network systems.

Hummingbird has three main components: Pipeline Parser, Optimizer and the Tensor DAG Compiler. For a given predictive model with given parameters and output dependencies this parser generates an Intermediate Representation (IR) object. The optimizer then selects an ideal implementation strategy for the the IR object. The final component, the tensor DAG compiler picks the IR object and integrates to tensor operations which leads to enhanced computations.

D. Evaluation Metric

In this project, we have chosen to measure the performance of models by measuring the accuracy of its predictions. For classification models, accuracy is the metric predominantly used to evaluate the performance of the model. Here, we define accuracy as

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

IMPLEMENTATION

A. Dataset Description

This data contains information about traffic accidents across 49 states of the United States. The timeline of data collected in this dataset is from February 2016 to December 2020. Multiple APIs are used to capture data from entities that broadcast traffic data such as US State Department of Transportation, Law Enforcement Agencies, traffic cameras and sensors on the road network [13].

The dataset contains 47 columns and currently has about 3 million entries. It has mainly 4 different kinds of data categories: traffic events; weather events; address events; points-of-interest events. Traffic events mainly contain information about the location of the event and description of the type of the event occurred (e.g: accident, broken vehicle, congestion, lane blocked .,etc). Weather events describe the various features of the weather during the accident(e.g: temperature, wind-speed, visibility ,.etc). Address events describe the location of the event occurrence from a street level to a city and state level. Finally, the points-of-interest category of data has important contextual information about the location(e.g junction, railway crossing, traffic signals ,.etc).

In this project, we use an attribute describing the *Severity* of the accident. This attribute is a categorical column having 4 categories from Level 1 to Level 4, where level 1 signifies least severe accident and level 4 being a fatal accident.

B. Software Requirements

Computing Platform: If the model is run on a local system the follow specifications are recommended: 12GB RAM, Windows/Mac OS, 6GB graphic card.

We recommend using Python 3.x and install the following to run the code:

```
pip install tensorflow
pip install keras
pip install scikit-learn
pip install — upgrade neural_structured_learning
pip install torch
pip install hummingbird-ml
```

Code Repository Link:
https://github.com/1997anirudh/ANC-Project-

C. Data Cleaning

The raw data in our dataset must be cleaned in order to observe significant patterns to get insights from the data. We detect and correct corrupt or inaccurate records by replacing or deleting them. We handle null values by imputing appropriate values (impute the mean in case of numerical features and most common value in case of nominal features), so that we do not have data loss by deleting null entries.

D. Exploratory Data Analysis & Illustrations

Exploratory data analysis was then carried out to discover patterns, test hypothesis and identify relations between different attributes of our dataset. This was achieved by using statistic and various graphical representations.

The following are some of our observations:

i. From the plot, we see that seven of the top ten states where the accidents occur are coastal states with California experiencing significantly higher number of accidents in comparison to others.

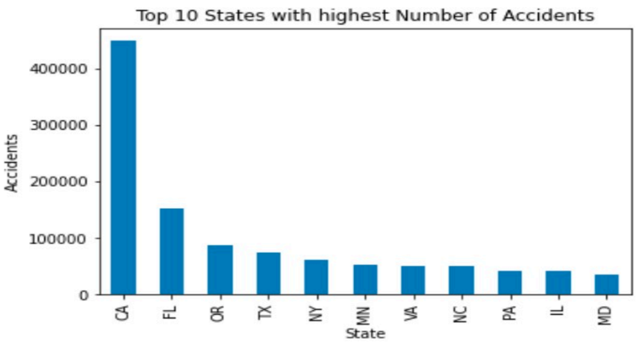


Fig 3: Plot of top 10 states with the highest number of accidents

ii.We can see from the below plot that there are fewer accidents on the weekends but the severity of accidents

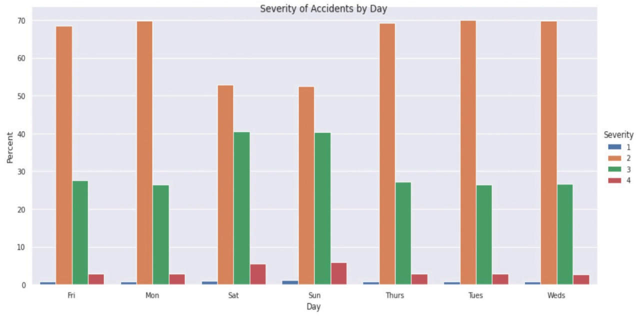


Fig 4: Distribution of accident severity by days of the week

increases. During the weekdays, level 4 type of accidents is under 5% of the total, but on the weekends we see that this percentage nearly doubles.

For the data cleaning and exploratory analysis the detailed code implementation and simulation can be found on the *EDA.ipynb* in the code-repository link mentioned above.

E. Feature Engineering

Before implementing the model, it is necessary to perform Feature Engineering to select the most ideal features suitable to our model. The process included principle component analysis to examine the correlation among different attributes and reduce information loss. We also used the standard scalar technique to fit the numerical features of data within a range and performed normalization.

F. Baseline Models

In this project, we chose Logistic Regression (LR), Decision Trees (DTree), Random Forests (RF), K-Nearest Neighbors (KNN) and MLP Classifiers (MLP) as baseline models. The baselines models are fed the train dataset with the features selected from the previous step. These models then predict the 'Severity' of accidents based on the features.

For logistic regression, we tested different solvers to find that the '*lbfgs*' solver gave the optimal output. We run the decision tree model over multiple iterations by varying the maximum depth of the model. In the random forest model, we iterate the model while changing the number of trees parameter from 100 to 200 to find the optimal output in our case. For the KNN model, we tuned the hyper parameter to optimize the number of neighbors and therefore the model output in this case. Finally for the MLP Classifier, we run the model and determine the ideal number of layers and neurons in each layers and also tune the learning rate to obtain the ideal output.

To implement and measure the performance of the baseline models, the *Baseline_models.ipynb* notebook can be executed from the code repository link given above.

G. Novel Machine Learning Models/Tools

In this project, we have enhanced upon the research work further by incorporating two new state of the art ML model/ tool.

i. Neural Structured Learning

Neural Structured Learning (NSL) is a novel ML model that can train neural networks and enhance the model to make far better predictions. Due to the vast size of this dataset, it is bound to have multiple points where erroneous data could be fed to it. We implemented an adversarial regularization wrapper function for our neural network model. When there are input perturbations that might confuse the model to give erroneous predictions, the NSL stands out as it identifies these adversarial inputs as implicit structured signals, learns from them and categorizes the points of interest into the appropriate severity levels and thus makes accurate predictions. For example, if the traffic sensor records data incorrectly or human errors occur while reporting an accident, the NSL model understands the issue by comparing with other attributes and corrects the loss to still make the right predictions.

After following the previous steps of data cleaning and featuring engineering, we move on to the preparation of data for model building with keras and tensorflow modules. Next, we use sequential base model function to create the neural network layers and include the ReLU and Softmax activation functions. We then wrap the model with adversarial regularization function. Finally we train the model and then evaluate the performance on the test data. The code implementation of this can be found in the *Neural_Structured_Learning.ipynb* notebook of code repository.

ii. Hummingbird

Hummingbird is a new ML tool to optimize deep learning pipelines. This latest tool makes use of tensor computations for training, testing and deployment. Hummingbird helps in compiling ML pipelines in the form of tensor computations that help deal with neural network systems and also optimize the usage of CPU/GPUs and vastly improve their computation capabilities.

In the accident domain, as the range and complexity of data will keep increasing with more real time relevant contextual data becoming available, it is necessary to build models that would utilize the data in its entirety while maintaining its capability to run computations and make accurate predictions. Integrating hummingbird with our ML models optimized the level of computations and scaling down the runtime multifold while handling the dataset we used which contained around 3 million records.

To implement this, we first install pytorch library followed by the installation of hummingbird-ml library. We then make use of the convert function in hummingbird library which transforms the built model to pytorch framework and implements tensor computations. This code is implemented in the *Humming_Bird.ipynb* notebook of the code repository.

SIMULATION AND RESULTS

After implementing the proposed ML models using this dataset, NSL model achieved an accuracy of 81%. As we can see from the figure, the proposed NSL model performs better when compared to the baseline models.

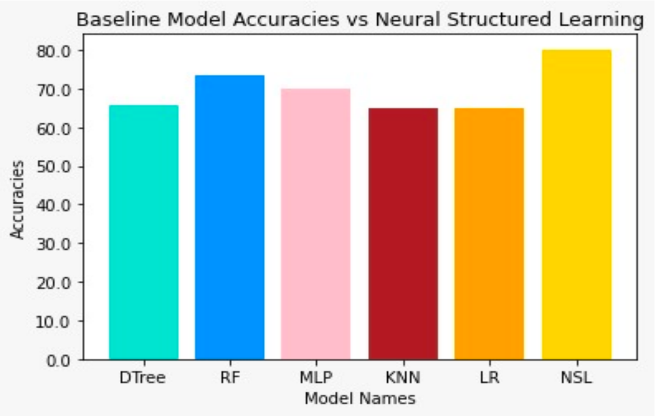
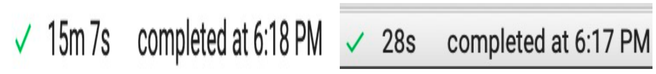


Fig 5: Accuracy of NSL model compared to other baseline models



✓ 15m 7s completed at 6:18 PM

✓ 28s completed at 6:17 PM

Fig 7: The running time before and after Hummingbird

Using the hummingbird tool, we were able to enhance the running time for example of a Random Forest model significantly as shown in the figure above.With Microsoft introducing this tool just a year ago, it is constantly working to integrate hummingbird with all machine learning models. In our project, as the dataset volume keeps increasing with time, Hummingbird would help a great deal in the future.

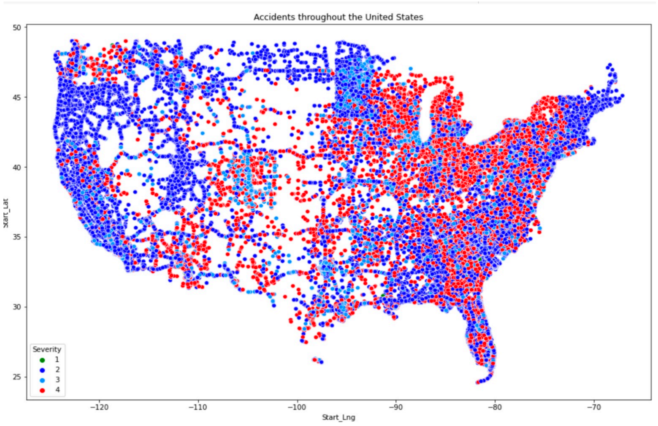


Fig 6: Accident Severity Distribution across the Country

As you can see, the majority of accidents are concentrated in the coastal areas of the country with severity level 2 and 3 being the most common accidents. We also observe that severity level 4 accidents follow along important highways and interstates in the countries.

DISCUSSION & WORK REMAINING

With the number of articles we see about accidents everyday and rise in autonomous vehicles clubbed with experiments on different modes of transportation we thought it would be ideal to pick the accident domain. We therefore picked this project [16] as it appealed to us and also had relevant social impact. This project uses a novel dataset that has been augmented and reinforced at a scale that has not been done before. As our first step, we wanted to enhance the project by building supervised ML models to predict the 'Severity' of accidents. The enhancement proposed from the paper was to use the entire dataset which consisted of approximately 3 million records to predict the *severity* of accidents compared to the paper's implementation which built a model for a subset of the data to predict the occurrence of accidents. While the project idea and domain were appreciated we received feedback after the initial review that novel ML models be included as part of the project. After discussing with the Professor to get guidance on how to proceed further, we decided to implement two latest ML model/tools that would further develop this project. both in terms of computations and model accuracies of the model's predictions turned out much better than the previously proposed models. Pending work remaining for the final report include code cleaning and annotations, ppt - presentation.

CONCLUSION & FUTURE WORK

This project implements supervised ML models using a country wide dataset to accurately predict the risk of accidents. This information can be provided to ground response teams like police forces, emergency response teams to be prepared for appropriate probable adversities. Also, urban planners can utilize this data to identify potential frequent points of accident and design road networks appropriately. In the future as the amount of data collection increases we would be able to obtain more important contextual information in real time which would aid in building models that predict accident severity accurately.

REFERENCES

[1] CDC, 'Road Traffic Injuries and Deaths—A Global Problem', 2020. [Online]. Available: https://www.cdc.gov/injury/features/global-road-safety/index.html

[2] CDC, 'Cost Data and Prevention Policies',2020. [Online]. Available: https://www.cdc.gov/transportationsafety/costs/index.html

[3] NHDSA, '2020 Fatality Data Show Increased Traffic Fatalities During Pandemic', 2021. [Online]. Available: https://www.nhtsa.gov/press-releases/2020-fatality-data-show-increased-traffic-fatalities-during-pandemic

[4] Ciro Caliendo, Maurizio Guida, and Alessandra Parisi. 2007. A crash-prediction model for multilane roads. Accident Analysis & Prevention 39, 4 (2007), 657–670.

[5] Li-Yen Chang. 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Safety science 43, 8 (2005), 541–557.

[6] Li-Yen Chang and Wen-Chieh Chen. 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of safety research 36, 4 (2005)

[7] Alameen Najjar, ShunâĂŹichi Kaneko, and Yoshikazu Miyanaga. 2017. Combining satellite imagery and open data to map road safety. In Thirty-First AAAI Conference on Artificial Intelligence. AAAI, Palo Alto, CA, USA.

[8]Zhuoning Yuan, Xun Zhou, and Tianbao Yang. 2018. Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatiotemporal data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM,

[9] Zhuoning Yuan, Xun Zhou, Tianbao Yang, James Tamerius, and Ricardo Mantilla. 2017. Predicting traffic accidents through heterogeneous urban data: Computing (UrbComp2017), Halifax, NS, Canada, Vol. 14. ACM, NY,

[10] Thang D. Bui, Sujith Ravi, and Vivek Ramavajjala. 2018. Neural Graph Learning: Training Neural Networks Using Graphs. WSDM 2018, Marina Del Rey, CA, USA bbd774a3c6f13f05bf754e09aa45e7aa6faa08a8.pdf

[11] TensorFlow, 'Neural Structured Learning', 2021. [Online]. Available: https://www.tensorflow.org/neural_structured_learning/framework#why_use_nsl

[12] 'Standardizing Traditional Machine Learning pipelines to Tensor Computation using Hummingbird', 2020. [Online]. Available: https://towardsdatascience.com/standardizing-traditional-machine-learning-pipelines-to-tensor-computation-using-hummingbird-7a0b3168670

[13] US-Accidents: A Countrywide Traffic Accident Dataset, 2021. [Online]. Available: https://smoosavi.org/datasets/us_accidents

[14] S. Moosavi et al. 2019'Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights'. SIGSPATIAL '19, November 5–8, 2019, Chicago, IL, USA

every team member is required to complete this form, which will be reviewed and signed by other team members								
by signing this form, you agree that the student 1) has contributed sufficiently to the team work, and 2) what is claimed by the student is accurate.								
attach this form at the end of your work report for each and every assignment. This form does not contribute to the page count.								
Assignment name	Progress Check 2							
Team #	6							
Student name Anirudh Srinivasan	worked on literature	worked on implementation (data, platform, test run, debug, compatibility...)	generated results (run results, result data processing, presenting results	wrote report (Intro, method, result, discussions, ...)	other significant contributions	peer approval 1	peer approval 2	peer approval 3
specific & detailed evidence is required to support claims of contributions (make reference to equation #, figure #, paragraphs, sections, etc...)	Studied and analyzed "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights" paper. Provided deeper insight into accident dataset to enumerate and determine the model building process. Worked on understanding the theory behind neural structured learning in the report	Set up the environment, included necessary libraries and modules while checking code compatibility and computation issues	Worked on building the neural structured learning model and integrating hummingbird model in the project. Helped in data processing for ML pipeline	Documented Methods and Abstract sections.	Upon discussions with TA and Professor researched upon novel ML models and tools namely NSL and hummingbird	Manohar	Pranav	Srikanth
Student name: Manohar Akula	worked on literature	worked on implementation (data, platform, test run, debug, compatibility...)	generated results (run results, result data processing, presenting results	wrote report (Intro, method, result, discussions, ...)	other significant contributions	peer approval 1	peer approval 2	peer approval 3
specific & detailed evidence is required to support claims of contributions (make reference to equation #, figure #, paragraphs, sections, etc...)	Studied and analyzed "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights" paper. Provided deeper insight into accident dataset by to enumerate and determine the model building process.	Set up the environment, included necessary libraries and modules while checking code compatibility and computation issues	Helped in enhancing model performance over existing supervised machine learning models	Helped in articulating the report in IEEE format and implementing citations in references section	Upon discussions with TA and Professor researched upon novel ML models and tools namely NSL and hummingbird	Anirudh	Pranav	Srikanth
Student name: Pranav Murali	worked on literature	worked on implementation (data, platform, test run, debug, compatibility...)	generated results (run results, result data processing, presenting results	wrote report (Intro, method, result, discussions, ...)	other significant contributions	peer approval 1	peer approval 2	peer approval 3
specific & detailed evidence is required to support claims of contributions (make reference to equation #, figure #, paragraphs, sections, etc...)	Studied and analyzed "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights". Worked on statistical analysis to determine performance indicators for model building process. Worked on understanding theory behind hummingbird in the report	Set up the environment, included necessary libraries and modules while checking code compatibility and computation issues	Worked on building the neural structured learning model and integrating hummingbird in the project. Helped in data processing for ML pipeline	Implemented Methods and Discussions section of the report	Upon discussions with TA and Professor researched upon novel ML models and tools namely NSL and hummingbird	Anirudh	Manohar	Srikanth
Student name: Srikanth Sampath	worked on literature	worked on implementation (data, platform, test run, debug, compatibility...)	generated results (run results, result data processing, presenting results	wrote report (Intro, method, result, discussions, ...)	other significant contributions	peer approval 1	peer approval 2	peer approval 3
specific & detailed evidence is required to support claims of contributions (make reference to equation #, figure #, paragraphs, sections, etc...)	Studied and analyzed "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights". Worked on statistical analysis to determine performance indicators for model building process. Worked on understanding the architecture of hummingbird.	Set up the environment, included necessary libraries and modules while checking code compatibility and computation issues	Worked on building the neural structured learning model and integrating hummingbird in the project. Worked on hyper parameter tuning for NSL	Worked on Implementation and Conclusion section of the report.	Upon discussions with TA and Professor researched upon novel ML models and tools namely NSL and hummingbird	Anirudh	Manohar	Pranav