Predicting Road Accident Risk Using Geospatial Data and Machine Learning (Demo Paper)

Yunzhi Shi
Amazon Web Services
shiyunzh@amazon.com

Raj Biswas
Amazon Web Services
rajbisr@amazon.com

Mehdi Noori
Amazon Web Services
nmehdi@amazon.com

Michael Kilberry
Amazon Web Services
kilberry@amazon.com

John Oram
HERE Technologies
john.oram@here.com

Joe Mays
HERE Technologies
joe.mays@here.com

Sachin Kharude
HERE Technologies
sachin.kharude@here.com

Dinesh Rao
HERE Technologies
dinesh.rao@here.com

Xin Chen
Amazon Web Services
xcaa@amazon.com

ABSTRACT

Over 100 fatalities and more than 8000 injuries are reported on average every day in the US caused by motor vehicle accidents. In order to provide drivers a safer travel plan, we present a machine learning powered risk profiler for road segments using geo-spatial data. We built an end-to-end pipeline to extract static road features from map data and combined them with other data such as weather and traffic patterns. Our approach proposes novel methods for data pre-processing and feature engineering using statistical and clustering methods. Our model achieves significant performance improvement for risk prediction using hyper-parameter optimization (HPO) and the open source AutoGluon library to optimize the ML model. Finally, an end-user visualization interface is developed in the form of interactive maps. The results indicate 31% improvement in model performance compared to baseline when model is applied to a new geo location. We tested this approach on six major cities in the US. The findings of this research will provide users a tool to quantitatively assess accident risk at road segment level.

CCS CONCEPTS

• **Computing methodologies** → Supervised learning by classification • **Applied computing** → Transportation

KEYWORDS

Traffic accidents, Machine Learning, Prediction, Classification

ACM Reference format:
Y. Shi, R. Biswas, M. Noori, M. Kilberry, J. Oram, J. Mays, S. Kharude, D. Rao, and X. Chen, 2021. Predicting Road Accident Risk Using Geospatial Data and Machine Learning (Demo Paper). In *SIGSPATIAL '21: ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, November 02–05, 2021, Beijing, China.*, ACM, New York, NY, USA. 4 pages. https://doi.org/10.1145/3474717.3484253

SIGSPATIAL '21, November 02–05, 2021, Beijing, China

ACM ISBN 978-1-4503-8664-7/21/11.
https://doi.org/10.1145/3474717.3484253

1 INTRODUCTION

Throughout the world, approximately 1.35 million people die each year as a result of road traffic accidents based on WHO statistics [1]. According to the U.S. Department of Transportation's Fatality Analysis Reporting System (FARS) [2], more than 8,000 people are injured, and over 100 people die in motor vehicle accidents on average every day in the United States. In this paper, we demonstrate a Risk Profiler for road segments using a Machine Learning (ML) approach which predicts accident risk in road segments on an hourly basis across any city, and hence enables safer route planning.

In the literature, there are a few variables considered for predicting an accident, including road and vehicle features, human factors and driving behavior, traffic flow and even vehicle trajectories. Yang et al. proposed a deep learning approach for real-time crash prediction on urban expressways. However, this approach is tested on a small set of data from only two expressways in Shanghai [3]. In addition, Rahim et al. developed a deep learning-based traffic crash severity prediction framework based on data at work zones in Louisiana from 2014 to 2018 [4]. Lei et al. utilized a tree-based algorithm and traffic flow continuity concepts to predict crashes in real time [5]. Moreover, Dong et al. proposed a hybrid approach for predicting car crashes in Knox County in Tennessee, which uses both supervised and un-supervised learning techniques [6]. Although most approaches proposed in the literature are promising, their scalability to new geo-locations are pending further testing. In addition, the explainability of these models are subject to question. Galatioto et al. presented a set of models for accident prediction in the UK, which are used for road safety simulation and predictions. Their result reveals high accuracy at national level in forecasting the number of casualties from a road crash and its severity [7]. However, they used granular parameters including average sidewalk width, and per cent of buses equipped with bicycle racks. This level of granularity makes it challenging to expand to new locations. In addition, researches have used deep learning approaches for leveraging the spatial and temporal dependency to predict fine-grained accidents

[8,9,11,12,13]. There are also applications of satellite imaging for mapping road safety [10]. However, these approaches may create computational and data dependency bottlenecks for implementing in real-world large-scale applications with near real-time latency capabilities.

The approach presented in this demo paper aims to fill the current gap in the research by unlocking the power of geo-spatial data in predicting car accident risk factor for road segments. Additionally, the feature engineering techniques used in this paper pave the path for a better model generalization and enable scaling up to a new geo-location. In this work, we do not intend to model crash severity, or make any conclusions on whether a particular road type is riskier than the other. Our conclusions are real-world data driven. This work provides a demonstration of an end-to-end reusable machine learning workflow using geospatial data on the cloud that can handle data acquisition, data processing, feature preparation, model training, model inference, and interactive visualization.

2 DATASETS

To determine the accident risk, the ML model needs to learn from relevant environmental data. These environmental factors include two types of data: static features that are attached to a geospatial location such as road geometry and signs; and spatial-temporal features that change over time such as weather and accidents. All datasets in this work are real-world data provided by HERE Technologies.

Static road features were collected from the map data for each road segment across the studied cities. Our data includes features such as road orientation, sinuosity, calibrated speed limit, whether the segment is a one-way road, whether the segment is at intersection, types of pavements, and traffic signs. Although there are lots of possible features that can be added to the list, we select these features with scalability in mind: the data should be universally available throughout different geographical regions so that ML models can learn to generalize across geolocations.

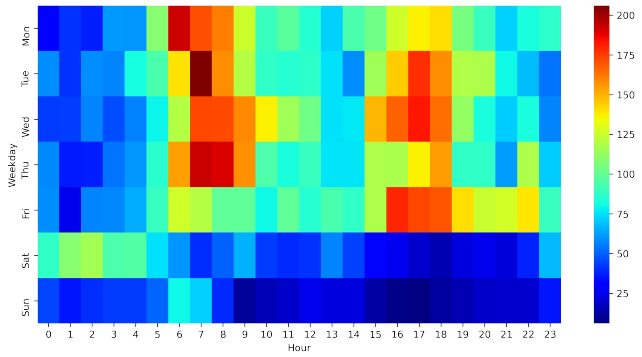

Figure 1: Chicago road accident heatmap per hour by day of the week (2018 – 2019).

We obtained historical accident records since 2018, and merged the weather information for the corresponding time and location where the accident occurred. The weather features include whether it was raining or snowing, visibility level (fog), temperature, humidity, and air pressure. The spatial location where each accident occurred is key to joining the spatial-temporal dataset with the static road features. By merging these two datasets, we built a table that lists the road and weather info for each historical accident event. This table becomes the positive samples for ML. Figure 1 shows the road accident pattern per hour by day of the week in Chicago during 2018–2019.

The negative samples can be synthetically generated based on the positive samples. Any other combination of road and weather conditions that is not in the positive sample dataset are negative samples i.e., no accident occurred on that road at that time. Using real-life accidental/non-accidental records will cause a significant imbalance between positive and negative examples. Therefore, we developed an algorithm to generate a certain number of negative samples so that each road segment has a proper mix of positive samples and negative samples. The ratio between negative and positive samples is a hyper-parameter subject to tuning. This allowed us to represent the pattern of positive-negative conditions in a practically sized dataset to train a ML model. Figure 2 shows the histogram of number of accidents per road segment. After excluding the ones without any accidents, most road segments had one accident while some had more than a hundred accidents per year.

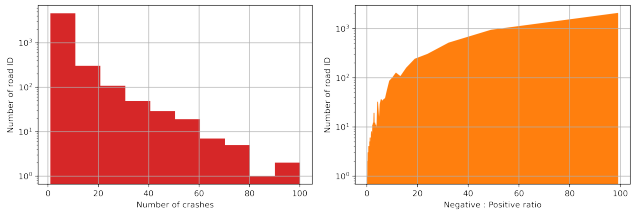

Figure 2: (Left) Number of accidents on each road segment per year, and (right) number of negative samples per positive samples on each road.

For each positive and negative samples, we computed additional spatial-temporal features including average traffic speed patterns, solar azimuth, and solar elevation. Figure 3 summarizes the workflow we used to prepare the dataset.

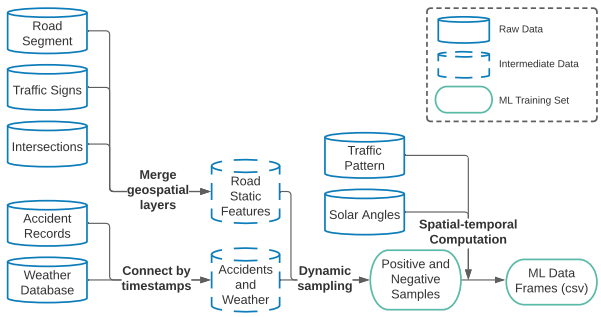

Figure 3: Summarizing the workflow of preparing geospatial data to ML training set.

Finally, the mix of positive and negative samples were split into three groups: train (65%), validation (15%), and test (20%). We

repeated this process for each of the six cities: Chicago, Columbus, Dallas, Miami, Salt Lake City, and San Francisco. In addition, we performed leave-one-out testing across six cities from where we collected the datasets. This testing mechanism was designed particularly to achieve highest generalization when the ML model is applied to a new geo-location without historical training data.

The training set compromises of 1.1M data points of road segments with static information, gathered from the aforementioned six metropolitan areas. We used around 50,000 accident events in total. All datasets are formatted in GeoJSON for further processing.

3 MODELING

We formulate the road accident risk prediction as a binary classification problem using tabular data, and use various feature data types as input, including Boolean/categorical values and floating-point number values.

3.1 Baseline model

The baseline model only uses raw data. We first train XGBoost models with the vanilla dataset directly from the preparation step to establish the baseline model, which is a benchmark reference to understand how the model improved in the next steps. Figure 4 shows quantitative scores of the model performances for the baseline scenario. We chose area under the curve (AUC-ROC) score to quantitatively measure the model performance instead of prediction accuracy. We observed that although all-city AUC score was 0.734, all leave-one-out test scores were less than 0.550, showing that the baseline model is incapable of generalization to a new city without that city's training data.

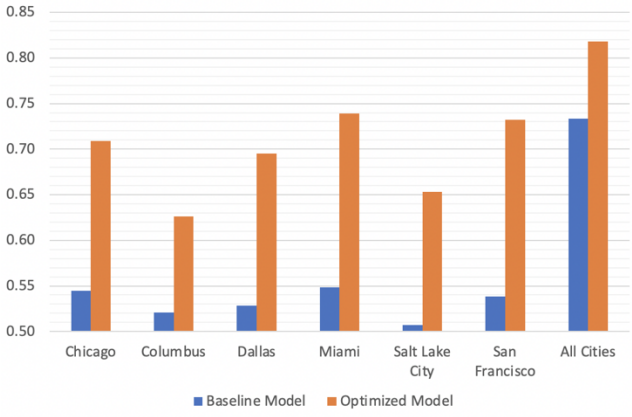

Figure 4: Performance comparison between scores (AUC) of the baseline model and optimized model after feature engineering.

3.2 Feature importance analysis

We would like our model to predict accident risks confidently across geographical regions. To understand the model better, we calculated the feature importance and shapely additive explanation (SHAP) value graphs to analyze how model reacts to all features in the training dataset [14]. Figure 5 shows the top 16 most important features for the baseline model. For example, the feature "One way" is considered to be the most important feature. The feature importance was used to guide feature engineering. For instance, among categorical features the unimportant categories were aggregated into a single category without affecting the model. We also observed that most of the road static features were prominent in the top important features and hence were grouped into one "Road cluster" feature to simplify the model.

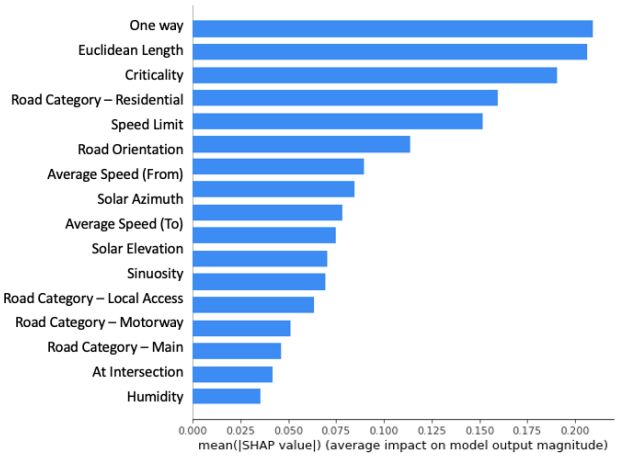

Figure 5: Feature importance analysis and SHAP values from the baseline model

3.3 Feature engineering and model tuning

Based on the previous analysis, feature engineering was designed to improve the learning efficiency. We identified closely relevant features that added unnecessary dimensions to the data. We also used heuristic knowledge to encode features by clustering, combined multiple features into a collective signal and proposed new features with real-world domain expertise. We introduced two important features that proved to significantly improve the prediction performance:

- Mean crash rate (MCR): The prior probability of an accident for each road segment with at least a minimum number of historical events.
- Static road clusters: using HDBSCAN algorithm to cluster all road static features into tokens. This method reduced the data dimension and exposed the significant component from a complex group of features.

We performed standard pre-processing on each feature based on their data distributions, including standard scaling and power transform. For instance, "Sinuosity" actually indicates the power transformed value of the sinuosity variable in the model's predictors. These are later used in subsequent ML training and analysis. The SHAP scatterplot in Figure 6 further elucidates the impact of feature importance. The scatterplot shows that a high value of a feature contributes positively to the model output whereas a lower value of this feature contributes negatively. It

can be seen that the new features such as the MCR and static road clusters contribute the most to the decision-making process.

In addition to feature engineering, techniques that optimize the model during the training process help improve the prediction performance. With the open-source AutoGluon library [15] deployed on Amazon SageMaker, we fit multiple ML models including XGBoost, random forest (RF), and neural networks (NN). It handled missing data and skewed data values, and selected algorithms automatically. Next, we used Bayesian optimization to select the best hyperparameter values for the chosen model. Eventually, the ML model achieves a substantial improvement compared to baseline: 30.51% and 11.52% in leave-one-out and all cities testing, respectively as seen in Table 1.

Table 1: Performance comparison of Baseline and Optimized models

Test Set	Baseline	Optimized	Improvement
Leave one out (average)	0.53	0.69	30.5%
All Cities	0.73	0.82	11.5%

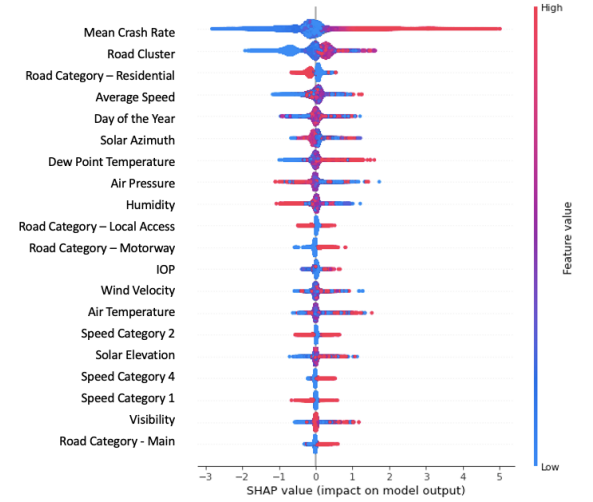

Figure 6: Feature importance analysis from the model after feature engineering.

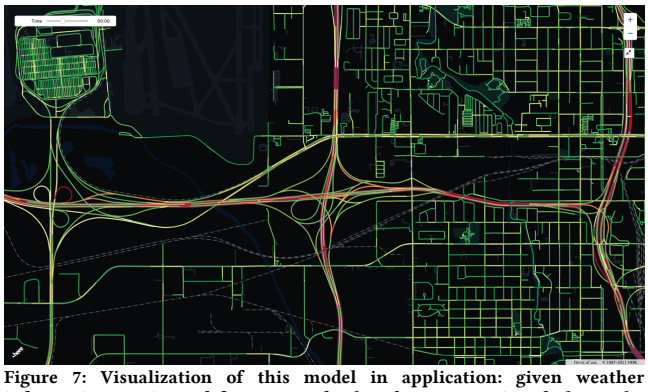

Figure 7: Visualization of this model in application: given weather information, static road features, and other data as mentioned above, the workflow can process and predict traffic risk scores on each road segments represented by colors: green-to-red means low-to-high traffic accident risk at the designated time.

We developed a visualization demo using the HERE map widget in the Jupyter notebook environment. In the visualization, color coding is selected to intuitively show accident risks: green as low, orange as medium, and red as high. The map widget allows basic interactive functions such as panning and zooming, and also advanced functions such as selecting time of the day or day of the week to visualize the estimated risk score map at any selected time.

4 DISCUSSIONS AND CONCLUSIONS

The model can be further improved by adding more historical data, more features such as billboard locations and more granularity of the weather data. The performance can also be improved by applying more feature engineering such as adding mean crash rate for static road clusters, for speed categories in each city and for hour categories in each city. Moreover, as stated in the literature review, deep learning approaches can be further explored if real-time predictions are not necessary for the application. The predictions of this model can be used to assess risk of roads by drivers, policy makers and transportation planners to identify safer routes for traveling. This paper demonstrates the end-to-end ML workflow unlocking the power of geospatial-temporal data through data preparation, feature engineering, model interpretation and model tuning. This work resulted in a machine learning model that can be scaled easily across geographic regions with an improvement of 30.5% over the baseline performance.

REFERENCES

[1] World Health Organization, 2020. Road traffic injuries, https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries/.

[2] U.S. Department of Transportation, 2021. Fatality analysis reporting system (FARS), https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars/.

[3] Yang, K., Wang, X., Quddus, M. and Yu, R., 2018. Deep learning for real-time crash prediction on urban expressways. *Transportation Research Board* (No. 18-04829).

[4] Rahim, M.A. and Hassan, H.M., 2021. A deep learning based traffic crash severity prediction framework. *Accident Analysis & Prevention, 154*, p.106090.

[5] Lei, T., Peng, J., Liu, X. and Luo, Q., 2021. Crash prediction on expressway incorporating traffic flow continuity parameters based on machine learning approach. *Journal of Advanced Transportation, 2021.*

[6] Dong, C., Shao, C., Li, J. and Xiong, Z., 2018. An improved deep learning model for traffic crash prediction. *Journal of Advanced Transportation, 2018.*

[7] Galatioto, F., Catalano, M., Shaikh, N., McCormick, E. and Johnston, R., 2018. Advanced accident prediction models and impacts assessment. *IET Intelligent Transport Systems, 12*(9), pp.1131-1141.

[8] Zhou, Zhengyang, et al. "RiskOracle: A Minute-Level Citywide Traffic Accident Forecasting Framework." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 01. 2020.

[9] Yuan et al. Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data, In KDD 2018.

[10] Najjar, Alameen, Shun'ichi Kaneko, and Yoshikazu Miyanaga. "Combining satellite imagery and open data to map road safety." Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[11] Zhou, Zhengyang. "Attention based stack res-net for citywide traffic accident prediction." 2019 20th IEEE International Conference on Mobile Data Management (MDM). IEEE, 2019.

[12] Huang, Chao, et al. "Deep dynamic fusion network for traffic accident forecasting." Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019.

[13] Zhou, Zhengyang, et al. "Foresee urban sparse traffic accidents: A spatiotemporal multi-granularity perspective." IEEE Transactions on Knowledge and Data Engineering (2020).

[14] Lundberg, S. and Lee, S.I., 2017. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

[15] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M. and Smola, A., 2020. AutoGluon-Tabular: robust and accurate AutoML for structured data. arXiv preprint arXiv:2003.06505.