WEATHER-RELATED TRAFFIC ACCIDENT PREDICTION USING RANDOM FOREST WITH CATBOOST AND LIGHTGBM

ABHILASHA GARG
MS-DATA SCIENCE

Research Proposal

AUGUST 2023

Abstract

Accurately forecasting traffic conditions during adverse weather events holds significant importance in optimizing transportation systems and mitigating potential risks. This study includes a thorough analysis of traffic accident prediction during weather-related events. The study harnesses the capabilities of two random forest algorithms: CatBoost, and LightGBM. The research focuses on using the Large-Scale Traffic and Weather (LSTW) dataset, a vast collection of real-world traffic and weather-related data for the United States.

Two main goals of this project are to estimate traffic patterns in the presence of weather disturbances using cutting-edge algorithms, and to meticulously compare and evaluate the predictive abilities of these algorithms. The initial phase encompasses data preprocessing of the LSTW dataset, entailing the extraction of pertinent features that capture the correlation between weather events and traffic accidents. Subsequent phases involve the rigorous training and evaluation of the CatBoost, and LightGBM models, aiming to get accurate predictions concerning traffic congestions.

Table of Contents

1. Background

Traffic accidents are strongly affected by weather-related factors. Notably, a thorough investigation by the National Highway Road Safety Administration (NHTSA) found that weather-related incidents accounted for 20% of all traffic deaths in the US in 2019. (Reish and Leah, n.d.)

These weather-related traffic fatalities are determined by a multitude of causes, including:

1. Inclement weather conditions such snow, rain, fog, and other meteorological phenomena can dramatically reduce visibility, finding it challenging for drivers to identify other cars and obstacles in the road.
2. The prevalence of icy road surfaces makes it harder for drivers to maintain control of their cars, enhancing the probability of accidents.
3. Violent wind gusts have the ability to cause vehicles to depart from their intended path, boosting the risk of accidents.
4. When roads are blocked subject to flooding, accidents can occur as a result of the vehicular manslaughter conditions.

For predicting traffic accidents, especially those affected by weather, machine learning has proven to be a powerful tool. Using machine learning to predict weather-related traffic accidents has become increasingly popular in recent years (Hasan et al., 2023; Luo and Wang, 2023; Testolina et al., 2023).

The LSTW dataset, a comprehensive collection of traffic and weather event data for the United States, is a noteworthy source in this area (Moosavi et al., 2019a). The training of models using random forest for forecasting weather-related traffic incidents can make good use of this dataset.

2. Problem Statement

Around the world, traffic accidents are a major source of fatalities and injuries. Each year, there are more than 6 million traffic accidents in the United States alone, resulting in more than 38,000 fatalities. Many of these collisions are brought on by particular sorts of traffic incidents,

such roadwork or severe weather. Various studies have been conducted by researchers (Moosavi et al., 2019b; Mondal et al., 2020; Hasan et al., 2023) across the globe with different dataset (Diaz-Ruiz et al., n.d.; Dadwal et al., 2021; Luo and Wang, 2023) in order to achieve accurate prediction based on various factors and techniques.

In this paper, we suggest developing a random forest algorithm model that could be used to evaluate big datasets of traffic data to find correlations that can be used to anticipate accidents. To create our models, we'll employ the two well-known algorithms LightGBM and CatBoost. Several criteria, including accuracy, precision, and recall, will be used to assess the performance. In addition to assessing how independent factors may affect the models' performance, we will also conduct sensitivity analysis and comparison between the two algorithms.

3. Research Questions

For each of the mentioned study objectives, the subsequent research questions are proposed.

1. Can accidents caused by certain weather conditions be predicted using random forest?
2. Determining a correlation between poor weather and road accidents.
3. What features are the most effective for projecting accidents driven on by climate or construction?
4. To review different techniques and evaluate the precision and model performance leveraging random forest algorithms.

4. Aim and Objectives

The primary goal of this study is to recommend the top model for using the random forest algorithm to forecast traffic accidents induced by meteorological occurrences in the United States.

The following are the study's objectives, which were based primarily on aims of this study:

1. To investigate the trend and correlation between unfavorable weather and traffic occurrences.
2. To determine the variables that are significant for forecasting traffic accidents.
3. To compare CatBoost and LightGBM predictive modelling
4. To assess the effectiveness of models using a range of metrics.

5. Significance of the Study

This study introduces an innovative approach using LightGBM and CatBoost algorithms for accurate accident prediction, focusing on feature selection and predictive model architecture. The integration of LightGBM and CatBoost is expected to improve accuracy and performance, especially in the Large-Scale Traffic and Weather (LSTW) dataset. The focus on reproducibility, reuse, and extensibility in the research community ensures transparency and reliability, facilitating future research endeavors in this domain.

6. Scope of the Study

For a number of reasons, the focus of this study is only on the forecasting of weather-related traffic incidents in the United States.

- First, there is a need for reliable ways to anticipate weather events because they are a significant contributing factor to traffic accidents.
- The LSTW dataset, a sizable set of weather and traffic occurrences for the United States, offers a thorough analysis of these occurrences.
- Third, two well-known machine learning algorithms, LightGBM and Catboost, may also be efficient in predicting traffic accidents.
- The final standard metrics for assessing the performance of machine learning models are accuracy, precision, and recall.

7. Research Methodology

The methodology used entails important procedures as the choice of target data, pre-processing the selected data, and many others. The following are these steps:

- **Data Collection**: The first step is to choose the target data. We will utilize the LSTW dataset, a sizable dataset of traffic and weather activities for the United States, in this research.
- **Data preprocessing**: Pre-processing the data is the subsequent stage. Data cleansing, outlier removal, and transforming the data into a structured and comprehensible format will be involved in this stage.
- **Data balancing**: The data may be imbalanced, meaning that there are more or fewer samples of one class than another. This can skew the results of the machine learning models. To address this, we will balance the data by oversampling or undersampling the minority class.
- **Feature Engineering**: The next step is to select the features that will be used to train the machine learning models. This is important because not all features are equally important. We will use a variety of techniques to select the features, such as univariate feature selection and recursive feature elimination.
- **Model training**: The next step is to train the machine learning models. We will use two popular machine learning algorithms, CatBoost and LightGBM, to train our models.
- **Model evaluation**: The final step is to evaluate the machine learning models. This includes evaluating the accuracy, precision, and recall of the models. We will also conduct sensitivity analysis to determine the impact of different factors on the accuracy of the models and compare the models of Catboost and LightGBM to find the best fit.

8. Requirements Resources

The following resources are necessary for this study:

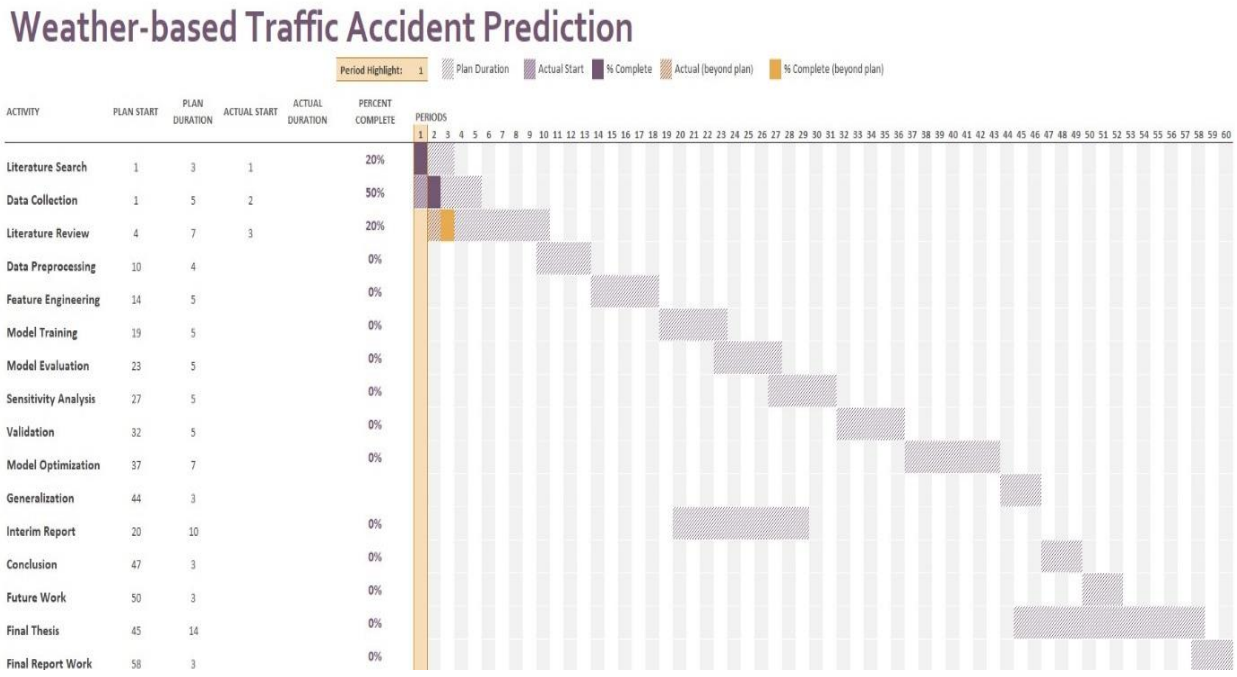
1. **Data**: The LSTW dataset is freely available.
2. **Software**: Data pre-processing, feature engineering, and machine learning will all require the usage of software. Python, R, and scikit-learn are some common choices.
3. **Hardware**: A computer with adequate memory and processing capability is required for this study.

9. Research Plan

Studies already conducted have shown that machine learning algorithms can be used to foresee mishaps that are caused by the weather (Moosavi et al., 2019c; Dadwal et al., 2021; Testolina et al., 2023; Wei et al., 2023). However, a thorough examination is necessary to fine-tune and prove their efficacy across various weather patterns and geographic areas of the United States.

The research plan for this study is as follows:

- Data collection and pre-processing
- Feature engineering
- Model training
- Model evaluation and validation
- Conclusion and future work



References

Dadwal, R., Funke, T. and Demidova, E., (2021) An Adaptive Clustering Approach for Accident Prediction. In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*. Institute of Electrical and Electronics Engineers Inc., pp.1405–1411.

Diaz-Ruiz, C.A., Xia, Y., You, Y., Nino, J., Chen, J., Monica, J., Chen, X., Luo, K., Wang, Y., Emond, M., Chao, W.-L., Hariharan, B., Weinberger, K.Q. and Campbell, M., (n.d.) *Ithaca365: Dataset and Driving Perception under Repeated and Challenging Weather Conditions*. [online] Available at: https://ithaca365.mae.cornell.edu/.

Hasan, A.S., Kabir, M.A. Bin, Jalayer, M. and Das, S., (2023) Severity modeling of work zone crashes in New Jersey using machine learning models. *Journal of Transportation Safety and Security*, 156, pp.604–635.

Luo, H. and Wang, F., (2023) A Simulation-Based Framework for Urban Traffic Accident Detection. Institute of Electrical and Electronics Engineers (IEEE), pp.1–5.

Mondal, A.R., Bhuiyan, M.A.E. and Yang, F., (2020) Advancement of weather-related crash prediction model using nonparametric machine learning algorithms. *SN Applied Sciences*, 28.

Moosavi, S., Samavatian, M.H., Nandi, A., Parthasarathy, S. and Ramnath, R., (2019a) Short and long-term pattern discovery over large-scale geo-spatiotemporal data. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, pp.2905–2913.

Moosavi, S., Samavatian, M.H., Parthasarathy, S. and Ramnath, R., (2019b) A Countrywide Traffic Accident Dataset. [online] Available at: http://arxiv.org/abs/1906.05409.

Moosavi, S., Samavatian, M.H., Parthasarathy, S., Teodorescu, R. and Ramnath, R., (2019c) Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In: *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*. Association for Computing Machinery, pp.33–42.

Reish and Leah, (n.d.) *Traffic Safety Facts 2019: A Compilation of Motor Vehicle Crash Data*. [online] Available at: https://crashstats.nhtsa.dot.gov/.

Testolina, P., Barbato, F., Michieli, U., Giordani, M., Zanuttigh, P. and Zorzi, M., (2023) SELMA: SEmantic Large-Scale Multimodal Acquisitions in Variable Weather, Daytime and Viewpoints. *IEEE Transactions on Intelligent Transportation Systems*.

Wei, W., Zou, S., Duan, W., Chen, Y., Li, S. and Zhou, Y., (2023) Spatiotemporal variability in extreme precipitation and associated large-scale climate mechanisms in Central Asia from 1950 to 2019. *Journal of Hydrology*, 620.