Machine Learning - CS4774
December 13, 2024

Accident Prediction Using VDOT Traffic Data

Team Members:
Varun Pavuloori - uja2wd
Nachiket Gusani - bya8tr
Avish Chiranth - wah6ne

Abstract

This project aims to develop a predictive model for traffic accidents in Virginia, more specifically in the state's capital Richmond, using traffic and weather data from the Virginia Department of Transportation (VDOT) and Traffic Records Electronic Data System (TREDS). By analyzing aggregated traffic metrics (e.g., Annual Average Daily Traffic), crash data, and monthly weather conditions (e.g., precipitation), we identify patterns contributing to accident likelihood. Our approach models the traditional pipeline of data pre-processing, exploratory data analysis, and testing different machine learning models to best predict total monthly accident counts. Preliminary results indicate that traffic volume and precipitation are significant predictors, with models such as Random Forest and K-Means Clustering performing robustly. The model, evaluated using metrics like mean squared error and $\mathbf{R^2}$, has potential applications in optimizing road safety interventions and resource allocation.

Motivation

Traffic accidents are a leading public safety concern, causing injuries, fatalities, and significant economic losses. In Virginia alone, more than 128,000 traffic crashes were reported in 2022, resulting in over 58,000 injuries and 1,000 fatalities, with an estimated annual economic impact of 3.7 billion dollars. These figures underscore the urgent need for targeted safety interventions to reduce the frequency and severity of accidents[1][3]. This project aims to help address this challenge by developing a machine learning model predicting the likelihood of traffic accidents using Virginia Department of Transportation (VDOT) data, including traffic volume, weather conditions, and related metrics. By identifying high-risk areas and periods, this model can inform road safety measures, optimize asset allotment for law enforcement and fire departments, and improve overall traffic management strategies. Additionally, the insights generated can support urban planners and policymakers in designing safer road networks and prioritizing infrastructure investments.

Method

For our predictive model, we have selected numerous machine learning algorithms including Random Forests, Logistic Regression, Support Vector Machines (SVMs), K-means Clustering, Gradient Boosting Regression, and XGBoost Regression to serve as classification tools. These algorithms are well-suited for handling classification tasks with mixed data types, such as those involved in traffic accident prediction. Logistic Regression provides a straightforward baseline to assess linear relationships between factors like traffic volume and accident likelihood, while Random Forests, SVMs, Clustering, and Boosting offer more flexibility in handling non-linear interactions, making them ideal for capturing complex patterns in the data. We chose these algorithms based on their successful application in related studies and their capacity to yield interpretable and accurate results.

In addition to the choice of models, we've dedicated significant effort to feature engineering. Traffic data, including AADT and the quality of the value, is combined with weather factors like precipitation and snow depth to provide a comprehensive input set against the total number of crashes on that road as reported by TREDS. By integrating weather data, we hope to capture the impact of environmental conditions on accident likelihood, particularly under hazardous situations like rain or snow. To reflect the actual implementation, the datasets were aggregated at the monthly level to align temporal features such as traffic volume, weather conditions, and crash counts. This multi-feature approach not only enhances the model's robustness but also allows us to experiment with the influence of each variable, aiming to isolate those with the most significant impact on accident probability. By predicting monthly accident counts rather than binary outcomes, this approach supports more granular insights to support the local community in traffic crashes. This layered approach to feature selection is expected to enhance the predictive capacity of our models.

Preliminary Experiments

Our preliminary experiments focused on creating a cohesive dataset, feature selection, and baseline model evaluation. We started by cleaning and merging the VDOT traffic data with NOAA weather data, addressing inconsistencies and handling missing values to ensure the models received proper inputs. Then, by taking the road alias of the route provided in the VDOT dataset and merging it with the crash reports from TREDS, we were able to complete a dataset with our desired features. This preprocessing stage was critical in reducing noise and improving the reliability of predictions. We also conducted initial exploratory data analysis to understand feature distributions, such as annual traffic volume and monthly weather conditions, which aligned with the temporal granularity of our aggregated dataset. This analysis helped us identify trends, such as increased crash counts during months with higher precipitation, and outliers that required further investigation.

For our initial model, we tested Logistic Regression to establish a baseline for comparison. Logistic Regression was chosen because it provides a simple yet interpretable framework for analyzing the linear relationships between features and accident likelihood. The model achieved very poor accuracy, which reflects variability in the data and highlights opportunities for improvement through more complex modeling techniques. Error analysis revealed that certain types of accidents, particularly those occurring in low-visibility conditions, were underpredicted. This insight highlighted the need for additional features to better capture the influence of weather and situational factors on accident likelihood. These findings informed our decision to explore more complex algorithms, such as Random Forests and Clustering, and experiment with hyper parameter tuning to enhance prediction accuracy and robustness. From this the hypothesis constructed was: traffic accidents can be accurately predicted using a combination of traffic volume and weather conditions, with non-linear machine learning models outperforming linear models in capturing complex relationships.

Intended Experiments

We originally planned to develop a classification model, yet pivoted to a regression model and began using algorithms including but not limited to Random Forests, Logistic Regression, and Support Vector Machines (SVMs) to predict the number of accidents. Additionally, we will explore clustering techniques to identify patterns in the data that may help group similar traffic conditions or accident hotspots. Our evaluation metrics will consist of mean squared error and R^2 between the actual vs predicted monthly accidents on a road segment. Furthermore, we will assess the model's performance across different years within Richmond Virginia, and incorporate cross-validation techniques to improve generalization.

The project will explore the impact of traffic volume, vehicle classification, and weather conditions on accident likelihood. The final model will provide a predictive framework that identifies high-risk road segments and time periods, allowing transportation authorities to take preventive measures with more updated datasets.

Results

We further compared the performances of a number of machine learning models by applying the criteria of Mean Squared Error (MSE) and R-squared (R^2). Among these initial models, the SVR had an MSE of 133.03 and an R^2 of 0.09, showing that it was hard to capture the underlying structure of the data. This might be because it is sensitive to hyper parameter settings, and out of the box, it has limited capability to model complex patterns without extensive tuning. Linear Regression, taken as the baseline model, did marginally better with its MSE coming to 126.88 and a very poor R^2 of 0.13 once again, showing how it is unable to handle such a dataset. Meanwhile, Random Forest Regression performed considerably better: it yielded an MSE of 60.75, with R^2 equal to 0.59, hence turning out to be highly effective, as this type of model has the capacity to capture non-linear relationships and feature interactions. Gradient Boosting Regression came in with an improved MSE of 54.61 and an R^2 of 0.63, while XGBoost Regression performed comparably, yielding an MSE of 61.17 and an R^2 of 0.58. Most notable, however, compared to the rest, was KMC with its top-of-pack ranking by a large margin, logging an extremely strong MSE of 47.20 and an R^2 of 0.68 in this case, proving very strong at modeling localized relationships.

After hyperparameter tuning, the models performed even better. Optimized KMC, with the best hyperparameters (`metric='manhattan'`, `n_neighbors=6`, and `weights='uniform'`), achieved an MSE of 45.19 and an R^2 of 0.69, making it one of the top-performing models. Gradient Boosting Regression, with the best hyper parameters (`learning_rate=0.3`, `max_depth=2`, `min_samples_leaf=2`, `min_samples_split=15`, and `n_estimators=250`), delivered the best overall performance with an MSE of 44.16 and an R^2 of 0.70. The tuned XGBoost model, using hyperparameters (`colsample_bytree=0.8`, `learning_rate=0.2`, `max_depth=3`, `n_estimators=200`, and `subsample=1.0`), performed closely behind Gradient Boosting with an MSE of 44.26 and an R^2 of 0.70.

Feature importance analysis marked precipitation as one of the most relevant features, emphasizing how weather conditions affect accidents. Precipitation is outranked in predictive importance only by measures of the quantity of road traffic, such as AADT and AAWDT (Annual Average Weekday Traffic). XGBoost and Gradient Boosting were among the top models since both iteratively improve the predictions, taking nonlinear interactions into consideration. Optimized KMC did a very good job by modeling local relationships despite being prone to be more sensitive with noisy high-dimensional data.

Interestingly, Linear Regression and SVR performed worse, pointing to the nonlinear relationship between most variables present in the data. Again, KMC did surprisingly well after tuning, given the fact that such a method is usually more prone to noise in data. One limitation of the present study concerns the grain of data: monthly aggregation may mask finer temporal patterns, like daily and hourly trends. Moreover, the model will be further limited in scope by the absence of variables such as road type or real-time conditions. In the light of these limitations, the current results also tend to underline the potential role that might be played by machine learning models in providing actionable insight aimed at improving road safety.

Contributions

This project was a collaborative effort among all team members, with each individual contributing to its success. Nachiket primarily focused on data gathering and cleaning, ensuring that the traffic, weather, and crash datasets were preprocessed and integrated effectively for analysis. Varun and Avish worked equally on developing, implementing, and evaluating the various machine learning algorithms used in the study. Additionally, the responsibilities for creating the presentation and video were shared equally among all team members, ensuring a cohesive and comprehensive delivery of the project results.

Discussions and Conclusion

Our study results again confirmed that machine learning models can be effectively used to predict the likelihood of traffic accidents based on data derived from traffic volume and weather variables. The best performances were recorded by Gradient Boosting and XGBoost, ensuring the highest R^2, thanks to their capability to handle a complex nonlinear interaction within the data set. The key predictor identified was precipitation, which corresponded to expectations that adverse weather conditions are an important factor in road accident risks. Similarly, the traffic metric features like AADT and AAWDT were very informative features, emphasizing that road usage is among the top determinants of accident probability. On the other hand, more baseline models such as Linear Regression and Support Vector Regression lagged behind, indicating that linear methods are not sufficient in this case to capture the rich relationships in the data. What these imply are important lessons from the interaction of traffic and weather conditions in predicting accidents.

The predictive model designed here could find many practical applications. It would enable transportation agencies and police to optimize resource allocation by increasing patrols or maintenance at times of hazardous weather in risky areas. Other useful insights, such as from urban planners and policy analysts, include giving priority to infrastructure improvements like improving drainage systems in those locations where heavy rainfall causes traffic jams consistently or redesigning dangerous intersections with high frequencies of accidents. Beyond immediate measures of safety, this approach should help formulate long-term traffic strategies to reduce injuries, fatalities, and financial losses that take place in road accidents.

Despite the promising results, some limitations of this study can still be noticed. Temporal patterns finer than a month could be obscured due to the work being based on monthly aggregated data, such as peak hours of traffic and seasonality of accident likelihood. Moreover, a host of critical features are missing in the dataset, which is necessary for model development to capture the comprehensive accident risk profile, such as road type, proximity to intersection, and real-time conditions of traffic flow. Another limitation may be related to the lack of driver-related factors: behavior, type of vehicle, or socio-demographic information that can affect the likelihood of an accident. These are limitations that show where the model needs improvements in its predictability.

Future research concerning this topic should focus on addressing the limitations identified in this study. Integrating more datasets, such as detailed road characteristics, real-time traffic, and driver behavior, could provide a clearer understanding of accident risks. Using more data, such as daily or hourly traffic and weather conditions, could enhance the model's temporal precision. Advanced methods like spatial clustering could help identify accident hotspots, while deep learning techniques might capture more intricate patterns in the data. Finally, extending the model to predict accident severity rather than occurrence could provide further value for decision-making in traffic safety.

References

[1] Virginia Department of Motor Vehicles. (2023). Virginia traffic crash facts: 2022. *Virginia Department of Motor Vehicles*. Retrieved from `https://www.dmv.virginia.gov/safety/crash_data/crash_facts/2022_crash_facts.pdf`

[2] Wang, J., Doe, J., & Smith, A. (2020). Traffic accident prediction using machine learning techniques. *Journal of Urban Transportation*, 42(1), 35–50. `https://doi.org/10.1007/s11067-020-09578-7`

[3] Heutel, G., & Miller, N. H. (2015). Weather, Traffic Accidents, and Climate Change. Resources for the Future. Retrieved from `https://media.rff.org/archive/files/sharepoint/WorkImages/Download/RFF-DP-15-19.pdf`

[4] Sackstein, I. (2020). How Does Weather Affect Traffic Accidents? Sackstein, Sackstein & Lee, LLP. Retrieved from `https://sacksteinlaw.com/weather-affect-traffic-accidents/`

[5] Bascom, R., & Meng, Q. (2016). Adverse Weather Conditions and Fatal Motor Vehicle Crashes in the United States, 1994–2012. Environmental Health, 15, 120. BioMed Central. Retrieved from `https://ehjournal.biomedcentral.com/articles/10.1186/s12940-016-0189-x`

Dataset References:

- VDOT Traffic Data:
  - `https://www.virginiadot.org/info/ct-TrafficCounts.asp`
- Virginia Weather Data:
  - `https://www.weather.gov/wrh/climate?wfo=akq`