Predictive Analysis of Crash Incidents in Hillsborough County

By: Bo Fethe

Group 123

March 2024

Table of Contents

1. Introduction

Hillsborough County, located in southwest Florida, is home to a diverse population and a rapidly expanding network of roadway. Given its urban communities, tourism appeal, access to natural resources, and well-developed transportation network, the county sees substantial Average Annual Daily Traffic (AADT) levels on its roadways. Unfortunately, along with this traffic comes the risk of road accidents and crashes, posing challenges to public safety and transportation management. Hillsborough County's population is rapidly growing and is expected to increase by 461,371 (+30.4%) from 2020-2050 (Plan Hillsborough, 2023), which causes roadway conditions to become more congested; and thus, increases the probability of traffic accidents occurring. The purpose of this report is to serve as a comprehensive examination of all crash data in Hillsborough County to offer valuable insights into the dynamics of traffic accidents to support safer roadways and communities.

Machine learning models trained on historical crash data can forecast the probability of future accidents along specific corridors or under certain conditions. Analyzing these events is crucial to recognize patterns, trends, and contributing factors, enabling authorities to prioritize safety improvements at specific locations. Many governments integrate principles from the Vision Zero initiative, aiming to eliminate fatalities and serious injuries in road traffic incidents, into their design strategies (Vision Zero Network, 2024). Predictive analytics enable stakeholders to incorporate best practices during the design phase, leveraging machine learning to comprehend spatial and temporal trends, pinpoint key crash contributors, and develop anticipatory models. By doing so, preventative measures can be integrated into transportation designs, mitigating the occurrence of crash events.

2. Methodology

2.1. Software

The project uses the programming language R version 4.3.2 ("Eye Holes") for statistical analysis. In addition, ArcGIS Pro version 2.9 is used for generating maps and ArcGIS Online is used for creating an interactive dashboard for data exploration.

2.2. Data Sources

Signal Four Analytics is a statewide geospatial crash analysis system developed and hosted by the University of Florida's GeoPlan Center (University of Florida, 2024). This system receives data from Florida's statutory custodian of records, the Florida Department of Highway Safety and Motor Vehicles (FLHSMV). The data used for this analysis was obtained on March 6, 2024, and filtered to all crash events in the year of 2023 within Hillsborough County.

2.3. Analytical Methods

This analysis uses machine learning to determine how well the severity of crash events can be predicted and if there are any regions with a statistically significant amount of crash events. Due to the nature of the data, it is anticipated the severity classification in the data raw data will be unbalanced as most crash events do not result in an injury, and that will need to be handled using techniques such as synthetic minority over-sampling technique (SMOTE) for the training datasets. In addition, the time values are placed into bins for the peak traffic hours of AM Peak (M-F; 07:00 – 10:00), Mid Peak (M-F; 10:00-16:00), PM Peak (M-F; 16:00-19:00), Weekend (Sa-Su; 06:00-20:00), and Off Peak for all other values (Florida Department of Transportation Systems Forecasting and Trends Office, 2023).

This analysis uses the following models to predict the severity of crash events occurring events occurring:

- Linear Discriminant Analysis (LDA): A classification model that finds the linear combinations of features that best define classes. Assumes normal distribution and predictor variables within each class have the same covariance matrix.
- Quadratic Discriminant Analysis (QDA): A classification model similar to LDA but allows each class to have its own covariance matrix.
- Naïve Bayes (NB): A probabilistic classification model based on Bayes' theorem. Assumes features are conditionally independent given the class label.
- Random Forest (RF): An ensemble bagging model that bootstrap aggregates multiple decision trees using randomly sampled subsets of data and features to predict or classify data.
- XGBoost (XGB): An ensemble gradient boosting model that iteratively builds decision trees with each iteration improving the previous tree and with an objective to minimize the loss function.
- Logistic Regression (Logit): A probabilistic classification model used for binary classification.

To measure performance, each model goes a series of Monte Carlo cross-validations and the mean square error (MSE) values are averaged.

Significant regions are determined using a spatial autocorrelation model, Getis-Ord Gi* (Gi*), which identifies clusters of high or low values based on the spatial relationships, physical distances, and the sum of neighboring attribute values, then spatially represents these clusters on a map. This model uses a distance-based weighting scheme where the weights increase as the distance decreases.

References

Florida Department of Transportation Systems Forecasting and Trends Office. (2023). *FDOT Source Book Methodologies: A Technical Report.*

Plan Hillsborough. (2023). *Hillsborough County attracting most new residents through 2050*. Retrieved from https://planhillsborough.org/hillsborough-county-attracting-most-new-residents-through-2050

University of Florida. (2024). *Florida Traffic Safety Dashboard*. Retrieved from Signal Four Analytics: https://signal4analytics.com

Vision Zero Network. (2024). *What is Vision Zero*. Retrieved from https://visionzeronetwork.org/about/what-is-vision-zero

Appendix