ML4VA: VA TRAFFIC CRASH HOTSPOTS FOR ROAD SAFETY AND MAINTENANCE PRIORITIZATION

Vaneesha Gupta
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA 22903
npw8uc@virginia.edu

Anjala Imam
School of Engineering and Applied Science
University of Virginia
Charlottesville, VA 22903
wxd4mw@virginia.edu

Emily Zhou
School of Arts and Sciences
University of Virginia
Charlottesville, VA 22903
csz6wd@virginia.edu

August 25, 2025

ABSTRACT

This project aims to enhance roadway safety by identifying roadways with a higher risk of vehicle collisions and prioritizing them for maintenance or safety interventions. A dataset of crashes occurring in Virginia was utilized to determine which road segments were more accident-prone based on factors such as roadway surface type, number of injuries, and crash severity. The project employed machine learning techniques, with a Random Forest Regressor used to predict crash severity as a continuous variable and a Random Forest Classifier to categorize roadway conditions into actionable categories. These predictions provide valuable insights into maintenance prioritization by identifying road segments associated with higher crash severity. Techniques such as target encoding, imputation, scaling, and feature engineering were utilized to preprocess the training data. The performance of the models was further enhanced through manual hyperparameter tuning and future steps include exploring advanced hyperparameter optimization techniques and incorporating additional data sources to improve the models' robustness and generalization.

1 Introduction

We aim to tackle the problem of deprecating roadway conditions and how they affect roadway safety, particularly in the aspect of accident prevention. We will be evaluating which roadways are in need of repairs by looking through historical traffic crash data in hopes of identifying high-risk roadways that have a greater number of occurrences and severity of accidents. Once we have identified these roadways, the scoring will provide insight into which roadways should receive funding in order to reduce the frequency and incidence of future crashes. The resulting model will serve as an application as we seek to produce actionable insight for local and state transportation agencies as well as policymakers to improve roadway safety as it is a leading cause of death in Virginia.

Some studies have been conducted in the past with a similar objective such as one described in https://ieeexplore.ieee.org/document/9418336. In this study, Khan et al. discuss Traffic Accident Analysis and Prediction Using Data Mining Techniques, and the prior methodology they used to investigate mainly focused on features such as the date and time of accidents, weather conditions, road conditions, vehicle types, and severity of the crash in order to determine the time of day and day of the week that had higher accident rates. Their process used clustering and then they applied other algorithms including SVMs to determine if it was a high or low-risk accident. They also discussed the importance of these features as well as visualization techniques to understand the problem.

Initially, we used a neural network to predict crash severity, but limitations such as overfitting and difficulty handling mixed data prompted a transition to Random Forest models. A Random Forest Regressor was employed for continuous crash severity prediction, while a Random Forest Classifier addressed roadway condition classification and repair prioritization. This dual-model approach improved accuracy, interpretability through feature importance, and robustness against mixed data types. Evaluation metrics such as precision, recall, F1 score, and RMSE helped confirmed the Random Forest models' effectiveness in addressing the dual objectives of crash analysis and infrastructure planning.

Dataset Used: The dataset we will use can be found at `https://dashboard.virginiadot.org/pages/safety/crashes.aspx`.

Contributions

Vaneesha Gupta worked on visualizing the data by providing insightful maps and correlation graphs and fixing issues in the data pre-processing steps, error analysis from training results, and conducted manual hyperparameter tuning.

Anjala Imam prepared and cleaned a complex dataset of over a million entries, conducting in-depth analysis to understand undocumented features and establish effective preprocessing steps for a solid data foundation and performed manual hyperparameter tuning.

Emily Zhou trained the baseline model on the data set, developed and implemented the multiclassification model, created the final prediction csv file and mappings that would be sent to the Virginia Department of Transportation, and conducted manual hyperparameter tuning.

2 Methods

The methodology for this study was designed to ensure accurate predictions of crash severity and actionable insights into roadway repair needs based on crash data from Northern Virginia. This section details the machine learning techniques applied and the preprocessing steps used to construct robust predictive models.

2.1 Data Collection and Preparation

The dataset was sourced from local crash records, filtered to include crashes occurring in Northern Virginia from the year 2021 onwards. This filtering ensured relevance to the study area and time frame.

Key preprocessing steps included:

- Handling missing data using `SimpleImputer` for numerical and categorical attributes.
- Standardizing numerical features with `StandardScaler` to normalize the data distribution.
- Encoding categorical features using one-hot encoding, integrated into a `ColumnTransformer`.
- Splitting the data into training and testing sets using an 80-20 split.

2.2 Modeling Approach

The study adopted a two-step modeling pipeline:

1. **Baseline Neural Network:** Initially, a simple feedforward neural network was implemented as the baseline model. While this approach captured non-linear relationships, it suffered from high variance and overfitting due to limited training data and feature complexity.
2. **Transition to Random Forest:** To address these challenges, we transitioned to Random Forest models, which are well-suited for mixed data types and provide better feature importance insights. Two separate Random Forest models were trained:
    - **Random Forest Regressor:** Used to predict crash severity as a continuous variable, allowing for more granular insights.
    - **Random Forest Classifier:** Used to classify roadway conditions and suggest repair priorities, addressing infrastructure-related objectives.

2.3 Pipeline Design

The preprocessing and modeling steps were combined into a unified pipeline, ensuring seamless integration of data transformations and model training. For multi-output learning, we employed the `MultiOutputClassifier` to handle simultaneous predictions for crash severity and roadway conditions.

2.4 Hyperparameter Tuning

Given the runtime limitations of traditional methods like grid search and randomized search, manual hyperparameter tuning was performed. This included iterative adjustments to:

- `n_estimators`: Number of trees in the Random Forest.
- `max_depth`: Maximum depth of each tree.
- `min_samples_split`: Minimum samples required to split an internal node.

2.5 Feature Engineering

Feature engineering involved deriving interactions between key attributes, such as weather conditions and roadway surface types, to enhance model interpretability. Additionally, minority class oversampling was applied to mitigate class imbalance in roadway condition classification.

2.6 Evaluation Metrics

Performance was evaluated using:

- **Regression:** Root Mean Squared Error (RMSE) for crash severity prediction.
- **Classification:** Accuracy, precision, recall, and F1-score for roadway condition classification.

2.7 Reproducibility

The codebase leverages Python and popular libraries such as `scikit-learn`, ensuring reproducibility. The pipeline is fully documented and can be replicated by following the provided scripts and data preparation guidelines.

GitHub Repository: ML4VA Crash Hotspot Detection

3 Experiments

This section outlines the experiments conducted to refine our crash severity prediction model and enhance its scope by introducing a secondary classification task for roadway repair prioritization. Key components include model evaluation, error analysis, and parameter optimization strategies.

We began with a baseline model using a Sequential Neural Network, chosen for its ability to capture non-linear relationships for crash severity prediction. While the model showed potential, it suffered from high variance and underperformed due to limited data and feature constraints.

To address the limitations of the neural network, we transitioned to Random Forest models. This choice mitigated overfitting, allowed for better interpretability through feature importance analysis, and handled mixed data types effectively. Initial tests confirmed that Random Forest outperformed the neural network, with improved metrics such as classification accuracy for severity prediction. Since the project also evolved to include a secondary classification task: predicting roadway repair needs. We developed a multi-output classification pipeline using `MultiOutputClassifier` to predict both crash severity and roadway conditions simultaneously. This approach allowed us to address the dual objectives of crash analysis and infrastructure improvement.

3.1 Hyperparameter Optimization

Traditional grid search and randomized search methods were tested for hyperparameter tuning but were abandoned due to significant runtime issues. Instead, we adopted a manual tuning approach, iteratively adjusting parameters like the `max_depth`, `min_samples_split`, `min_samples_leaf`, `max_features`, and `n_estimators` for each subtask (e.g., crash severity, roadway defects). Class weights were adjusted to improve performance for minority classes.This approach proved both efficient and effective in identifying an optimal configuration.

3.2 Error Analysis

Error analysis revealed common misclassifications in scenarios involving rare crash types or underrepresented roadway conditions. This insight informed our preprocessing steps, such as oversampling minority classes and engineering features to capture interactions between weather conditions and roadway attributes.

3.3 Future Directions

Further experiments could include:

1. Incorporating additional external datasets for broader generalization.
2. Exploring ensemble methods to further enhance predictive accuracy.
3. Refining feature selection to improve computational efficiency.

4 Results

The final Random Forest model achieved the following outcomes:

- **Crash Severity Prediction:** Weighted F1-score improved significantly compared to the baseline. RMSE was reduced to 0.9519 with tuned hyperparameters (`max_depth`=15, `n_estimators`=75).
- **Roadway Defects:** Performance improved for minority classes by increasing `min_samples_leaf` to 6 and adjusting `max_features` to `sqrt`. Class weight adjustments further enhanced precision and recall.
- **Light Condition and Roadway Surface Condition:** Combining rare classes and using balanced class weights improved average F1-scores, demonstrating better representation across all classes.

Visualizations of feature importance highlighted key predictors, such as `WEATHER_CONDITION` and `TOTAL_INJURIES`. Spatial heatmaps identified high-priority areas for intervention, providing actionable insights for policymakers.

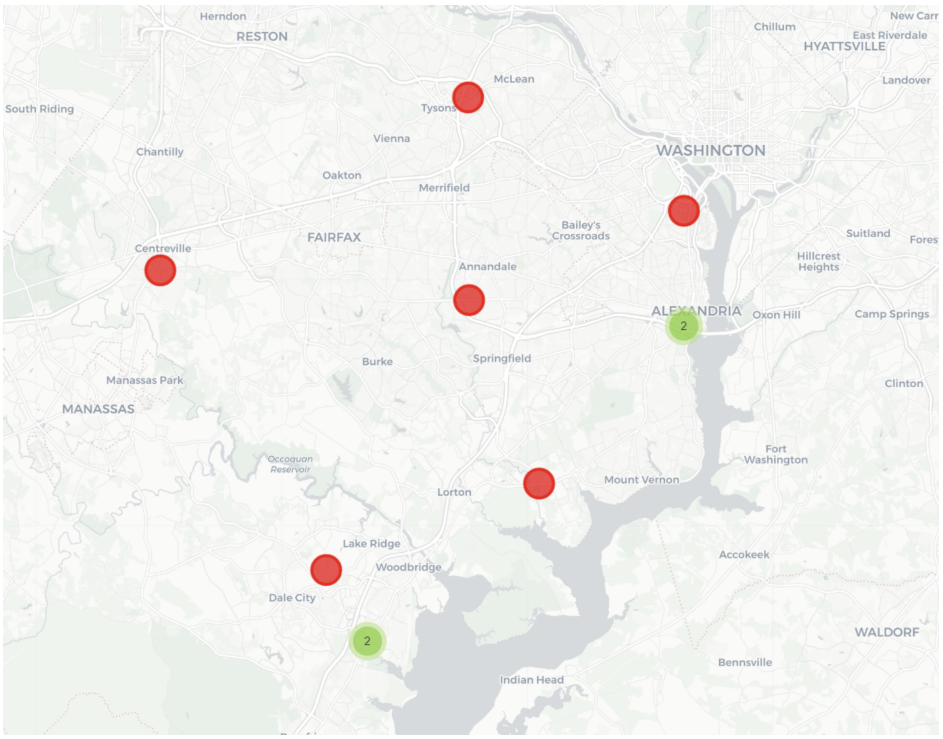
Summary of Key Metrics

Metric	Baseline Model	Final Model
Roadway Defect F1-Score	0.7233	0.9587
Light Condition F1-Score	0.0029	0.1092
Roadway Surface Condition F1-Score	0.0015	0.0333
Crash Severity RMSE	1.28	0.94

Table 1: Performance Comparison Between Baseline and Final Model

These results demonstrate the efficacy of the methodology and the improvements achieved through iterative experimentation and manual tuning. The RMSE score was utilized for regression analysis while the macro F1 score was calculated for the classification tasks. We focused on our F1 Score for our classification tasks due to the F1 Score being a harmonic mean of precision and recall which provided a balanced measure of a model's performance, especially with imbalanced datasets. The F1 score for Light Condition and Roadway Surface Condition was particularly low due to the extremely imbalanced datasets even with SMOTE applied.

- Crash Hotspot Prediction Notebook
  - `https://colab.research.google.com/drive/16S2w46Tb5U_gvGPDk_T3WYebFqypmn47?usp=sharing`
- Northern Virginia Hotspots Map



5 Conclusion

The results of this study demonstrate progress in addressing roadway safety concerns in Virginia by identifying accident-prone road segments and providing actionable insights for prioritizing maintenance and repair efforts. Through the application of machine learning techniques, particularly the transition to Random Forest models, we successfully predicted crash severity and roadway conditions, achieving improvements over the baseline neural network. The improved performance of the final model, as evidenced by the reduced RMSE for crash severity and the improved F1 scores for classification tasks, validates the hypothesis that data-driven approaches can effectively identify high-risk roads and support infrastructure decision making.

Our findings underscore the importance of key predictors, such as weather conditions and the number of injuries, in determining crash severity, although features like light condition and roadway surface condition may have limited predictive value and could be considered for removal or down-weighting in future models. Visualizations, including spatial heatmaps, further highlighted high-priority areas for intervention, offering a resource for state transportation agencies and policymakers to allocate resources efficiently. The ability to identify roadway segments associated with severe crashes has direct implications for the well-being of Virginia's residents, as targeted maintenance can reduce accident rates, improve overall road safety, and minimize economic and emotional burdens associated with vehicle collisions.

Despite these successes, the study encountered certain limitations. The complexity of multiclass classification posed challenges, particularly with minority class representation. While manual hyperparameter tuning and class weight adjustments improved performance, further optimization remains a potential area for exploration. Additionally, the scope of the dataset was limited to Northern Virginia, and the inclusion of external data sources, such as traffic flow, construction records, or real-time weather updates, could enhance model generalizability and robustness.

Future work should focus on refining feature engineering to include less influential attributes, exploring ensemble techniques for improved accuracy, and expanding the data set to include diverse geographical and temporal variations. Furthermore, advanced hyperparameter optimization methods, such as Bayesian optimization, could be explored to streamline the tuning process and achieve further performance gains.

Therefore, by identifying high-risk road segments and providing data-driven recommendations, this research lays the foundation for proactive, targeted interventions that can enhance the safety and well-being of Virginia's residents. As transportation agencies adopt and build upon these insights, the potential for safer roads and reduced accident rates becomes increasingly attainable.

References

[1] Brownlee, J. (2021, March 17). SMOTE for Imbalanced Classification with Python. MachineLearningMastery.com. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[2] Hyperparameter tuning. GeeksforGeeks. (2023, December 7). https://www.run.ai/guides/hyperparameter-tuning

[3] Koehrsen, W. (2018, January 9). "Hyperparameter Tuning the Random Forest in Python Using Scikit-Learn." https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

[4] Lee, E. (2023, April 11). Hyperparameter optimization: Grid search vs. Random Search vs. Bayesian Optimization in Action. Medium. https://drlee.io/hyperparameter-optimization-grid-search-vs-random-search-vs-bayesian-optimization-in-action-106f99b94e32

[5] Run.ai. (n.d.). Hyperparameter tuning. Hyperparameter Tuning: Examples and Top 5 Techniques. https://www.run.ai/guides/hyperparameter-tuning

[6] SVM Hyperparameter Tuning Using GridSearchCV | ML. GeeksforGeeks. (2023b, January 11). https://www.geeksforgeeks.org/svm-hyperparameter-tuning-using-gridsearchcv-ml/

[7] Virginia Department of Transportation. (2024, December 8). VDOT Dashboard: Crashes. https://dashboard.virginiadot.org/pages/safety/crashes.aspx