Introduction to Machine Learning

Final Paper

12/14/2024

Colin Bitz, Will Mitchell, Sam O'Brien, Lucas Vallarino

Executive Summary

This study examines the seasonal and geographical patterns of road accidents in Virginia to identify the most dangerous months for driving. Using machine learning techniques, the project provides data-driven insights to aid in travel planning. The analysis leveraged a dataset of accident occurrences, incorporating factors such as weather conditions, road conditions, location, and time. After an initial experimentation with K-Nearest Neighbors (KNN) for classification, a random forest regressor was implemented to predict monthly accident counts. This shift allowed for greater accuracy in capturing the relationship between variables and accident likelihood.

The model achieved impressive predictive performance, with a test R^2 of 0.9095, which showcases its capability to effectively predict upcoming traffic counts. Feature importance analysis revealed precipitation as the dominant predictor of accident counts, followed by road-specific and temporal factors. While these findings align with common-sense expectations as adverse weather conditions significantly impact driving safety, the model also predicts specific periods of time that experience an unproportionate amount of accidents.

Results indicate that November through January are the riskiest months for driving in Virginia– something that can be attributed to increased holiday travel and seasonal weather challenges such as snow and rain. July also emerged as a high-risk month, possibly due to summer travel spikes.

The insights gained are practical and actionable– providing drivers with information that suggests additional caution during holiday and summer travel seasons, especially during adverse weather conditions. While these results certainly benefit drivers, policymakers and transportation planners can use these findings to prioritize road safety measures during high-risk periods–

possibly planning for additional traffic, scheduling construction work, or ensuring that there are an appropriate number of emergency services.

Introduction

Road safety is a critical concern for everyone—from inexperienced teens and elderly drivers to policymakers, transportation planners, and the Department of Motor Vehicles. In Virginia, a state with diverse geography and all four distinct seasons, seasonal and regional variations significantly influence road safety risks. For instance, the icy and snow-covered roads of winter pose unique challenges, while summer brings a different set of risks related to increased travel activity, distracted driving, and road congestion. Understanding the primary factors that impact these patterns is crucial for mitigating risks, optimizing traffic management strategies, and ensuring an overall safer travel experience.

This study addresses the question: What are the worst months to drive in Virginia based on accident occurrences, and what are the underlying contributing factors? The motivation stems from the severe and wide-ranging consequences of road accidents, which impose not only economic burdens but also emotional and social costs. The National Highway Traffic Safety Administration (NHTSA) estimates that car crashes cost the U.S. economy hundreds of billions of dollars annually, including costs associated with medical care, property damage, and lost productivity(National Highway Traffic Safety Administration). Beyond economics, accidents impact the well-being of families and communities, highlighting the urgent need for targeted safety interventions. For policymakers and uniformed personnel, data-driven insights can enable better resource allocation, whether through improving winter road maintenance, scheduling construction projects, or deploying emergency services during peak accident periods. With advancements in machine learning, there is now an opportunity to leverage large datasets to identify high-risk conditions, locations, and periods with greater precision than traditional statistical methods would allow.

Initially, a supervised KNN classification model was proposed to classify roads into "High Risk," "Medium Risk," and "Low Risk" categories. However, this was not sufficient for the scope of the project as it failed to capture the variety of conditions that may impact the likelihood of an accident occurring. After further analysis, it was revealed that a random forest regressor was better suited for predicting accident counts. This shift was driven by the need for more nuanced predictions that accounted for a broader range of variables including but not limited to traffic volume, precipitation, and temporal trends. The dataset, comprising accident records across Virginia, provided insights into how environmental and temporal factors influence accident frequencies.

The findings reveal that November through January are consistently high-risk months, leading us to a probable belief that the holiday season in conjunction with increasingly poor weather conditions contribute largely to an increased risk of accidents. July also emerges as a high-risk month– possibly attributable to an increase in summer travel. The model's strong performance metrics (test R^2: 0.9095) validate its effectiveness for predicting accident counts, though its applicability to less-trafficked roads remains limited. This outcome makes sense, as harsher weather conditions make driving more difficult, and these months experience significant movement across the state. This includes college students returning home for the holidays, families traveling during summer vacations, and people visiting relatives during Thanksgiving.

This paper is organized into several sections to provide a clear and comprehensive analysis of road safety patterns in Virginia. The Data Section describes the dataset and preprocessing steps, including standardizing street names, handling missing values, and engineering features like temporal trends and precipitation levels to enhance model performance. Key challenges, such as defining a consistent risk metric and accounting for variations in

population density, are also discussed. The Methods Section explains the modeling approach, highlighting the choice of the Random Forest Regressor for its ability to capture non-linear relationships and assess feature importance. The data preprocessing pipeline, including one-hot encoding, normalization, and data splitting, is detailed alongside the evaluation metrics used to measure the model's accuracy, such as R^2 and Mean Absolute Error (MAE). In the Results Section, the model's predictions and findings are presented, revealing November through January as high-risk months due to holiday travel and adverse weather conditions, while July also shows elevated risk likely due to summer travel. Precipitation is identified as the most influential factor, underscoring its role in road safety. The strong model performance, with a test R^2 of 0.9095, validates its reliability in capturing these trends. Finally, the Conclusion Section reflects on the study's findings and their practical implications for drivers and policymakers. It highlights opportunities for further research, such as incorporating real-time traffic data and expanding the model's applicability to less-trafficked roads, to enhance road safety insights across Virginia.

Data

The dataset included accident occurrences in Virginia, with variables like location, date, weather, road type, and severity. Each observation represented a single accident. Preprocessing steps included one-hot encoding categorical variables (e.g., road names, months) and engineering features like precipitation levels to enhance model performance.

Key challenges included defining a consistent "risk" metric across diverse road types and addressing population density variations. Future iterations could incorporate datasets on traffic volume to better normalize accident rates by road usage.

Below are some of the processes which were conducted during the process of data wrangling– including standardization, grouping, and our approach to handling unknown, missing, or invalid data:

1. **Standardizing Street Names**
    a. The code uses a function to standardize street names, removing directional suffixes (e.g., "N," "S") and hyphens. This ensures consistency in the street data, reducing noise during analysis.
2. **Top 150 Streets**
    a. Accident data is grouped by street, and the top 150 streets with the highest accident counts are identified for analysis. This filtering step aligns with my earlier mention of focusing on high-traffic areas but emphasizes that this process is performed dynamically within the code.
3. **Handling Missing Data**
    a. Missing street names are filled with the placeholder "Unknown" before further processing. This ensures no null values interfere with analysis.
4. **Feature Engineering**
    a. Temporal trends and road-specific accident counts appear to be a focus, with streets filtered based on the identified high-risk category.

Methods

Objective and Model Choice

The objective of this study was to predict monthly accident counts on Virginia roads to identify high-risk periods and areas. After initial exploration, the Random Forest Regressor was selected as the primary model. This choice was motivated by its ability to handle non-linear relationships and assess feature importance effectively.

Data Preprocessing

1. **Standardization and Cleaning**
    a. Street names were standardized by removing directional suffixes and hyphens to ensure consistency across the dataset.
    b. Missing street values were replaced with "Unknown," ensuring completeness.
2. **Feature Selection**
    a. Temporal features, such as the month of the accident, and environmental features, like precipitation levels, were retained for analysis.
    b. Streets were ranked by accident counts, and only the top 150 streets were included in the analysis to focus on high-risk areas.
3. **Feature Engineering**
    a. Categorical variables like street names and months were one-hot encoded for compatibility with the regression model.
    b. Continuous variables like precipitation and wind speed were normalized to prevent feature scaling issues.
4. **Data Splitting**
    a. The dataset was split into training and testing sets using an 80-20 ratio to ensure robust evaluation of the model's performance.

Modeling and Evaluation

1. **Model Training**
    a. The Random Forest Regressor was trained on the preprocessed data. This ensemble method aggregates predictions from multiple decision trees, making it robust to overfitting and effective for capturing complex feature interactions.
2. **Hyperparameter Tuning**
    a. Default hyperparameters were used in this iteration, with opportunities for future tuning to optimize model performance.

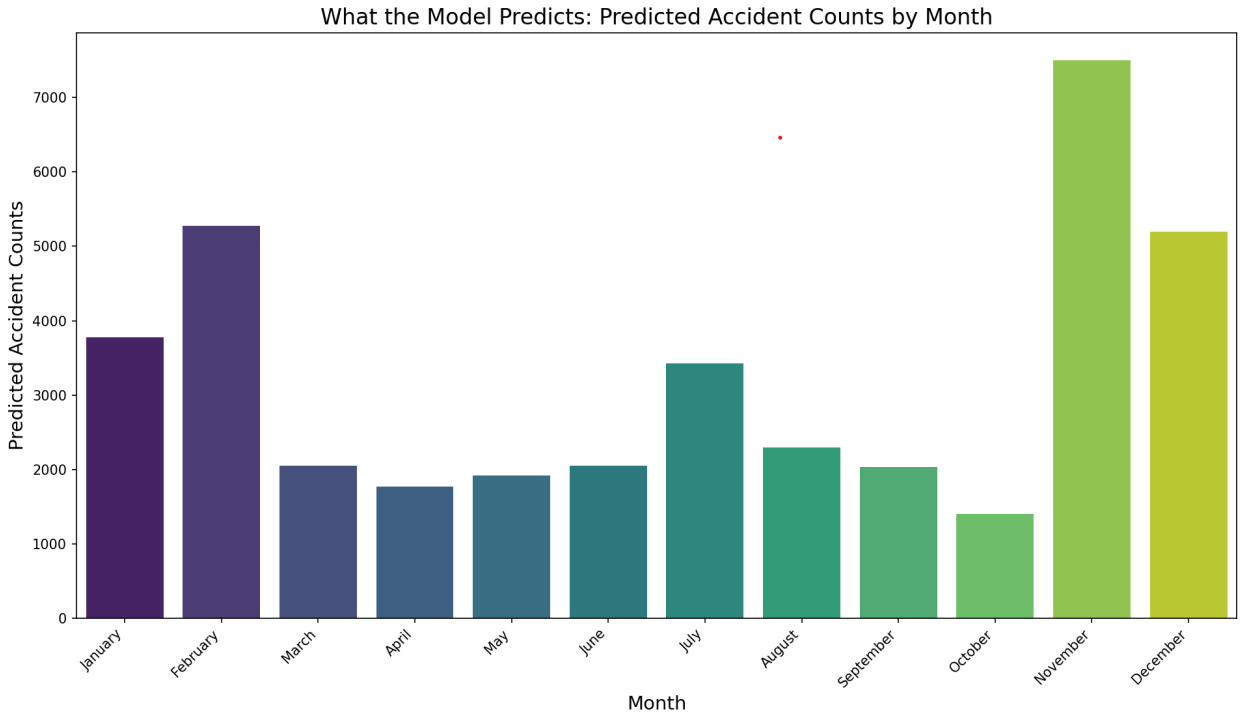
Results

Model Performance

The Random Forest Regressor demonstrated strong performance on both training and testing datasets:

- **Training R^2:** 0.9825
- **Testing R^2:** 0.9095
- **Training MAE:** 15.14
- **Testing MAE:** 39.64

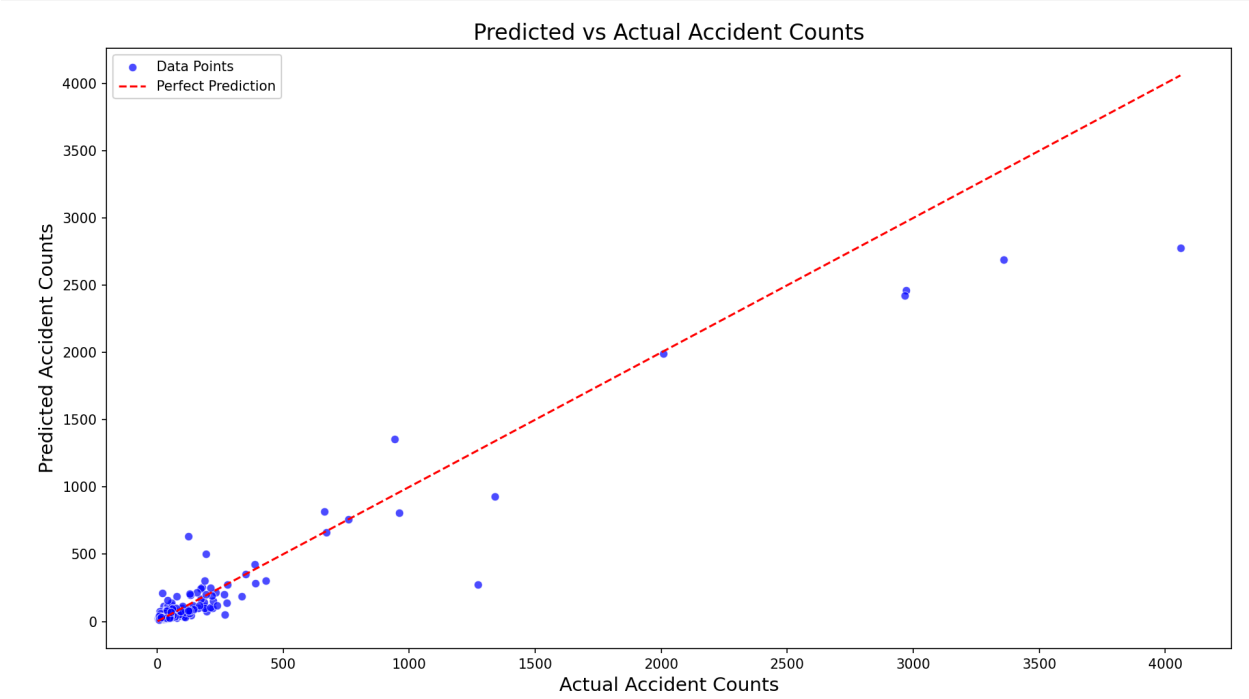
These results indicate that the model captures the underlying patterns in the data effectively while maintaining good generalizability.



The bar plot shows the model's predictions of accidents per month.

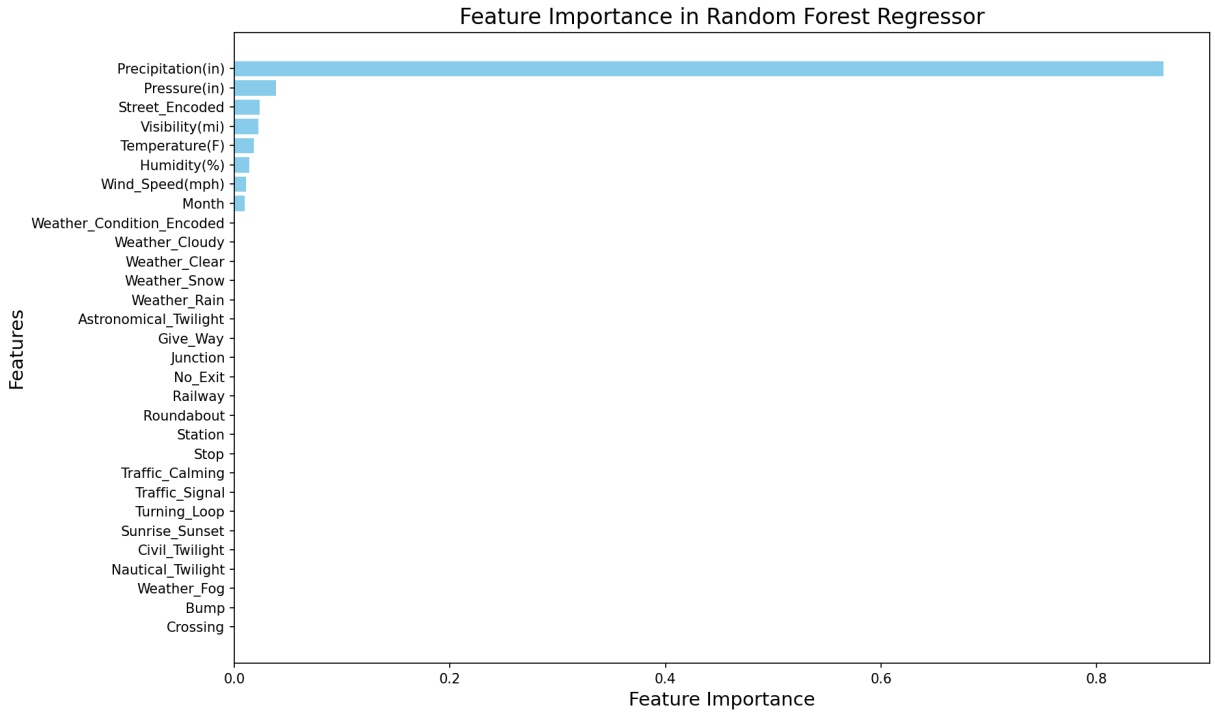
- The model's results follow common logic as the holiday seasons appear to be the worst months of the year to drive in Virginia.
- You can observe 2 tails at either end of the chart, demonstrating the increase of accidents
- The key takeaways from this graph and the overall model purpose are:
    - Holidays, specifically November through January, appear to be the worst times to drive in Virginia.

- Try also to avoid road-related travel in July, as there appears to be an increase in accidents during that specific month.



This graph demonstrates the model's performance. There are a few key takeaways to note:

- Most of the data points are clustered near the bottom-left where the values are close to 0
  - Indicator that there is a high concentration of months with lower accident counts in Virginia with less months having higher counts.
- The model shows a few outliers, which may explain why there are some instances of the model failing to predict instances of higher months more accurately.



This graph demonstrates what features the random forest model most often selects:

- The precipitation amount dominates the prediction of accident counts in this model. This is extremely logical as it is common knowledge that rain, snow, sleet, and other forms of weather make driving much more dangerous.
- The model uses other variables like street, and wind speed, and month, to help round out the decisions, but almost all of the prediction is associated with the precipitation amount.
    - This is useful because it suggests that precipitation amount is a heavy determinator of accident occurrences.

Key Findings

1. **Seasonal Risk Patterns**
    a. The analysis identified November through January as the riskiest months, primarily due to holiday travel and adverse weather. July also emerged as a high-risk period, potentially linked to increased summer travel.
2. **Feature Importance**
    a. Precipitation was the most influential predictor, reinforcing its critical role in road safety. Other important features included street identifiers and month, suggesting the combined impact of location and temporal factors.
3. **Visualization Insights**
    a. Visualizations of predicted versus actual accident counts highlighted the model's strengths in capturing low-accident months while revealing a few outliers for

high-accident months. These outliers suggest a need for additional data to improve predictions in such cases.

4. **Generalization to Major Roads**
    a. By focusing on the top 150 streets, the model effectively captured high-risk areas but lacked applicability to less-trafficked roads.

Conclusion

This study provides valuable insights into the patterns of road accidents in Virginia, addressing both seasonal and locational risks. Through the application of machine learning techniques, particularly the Random Forest Regressor, the analysis identifies November through January as the most dangerous months for driving, primarily due to holiday-related travel spikes and adverse weather conditions such as snow, sleet, and freezing rain or hail. The combination of increased road congestion during holidays and challenging winter weather significantly elevates. July also presents a higher-than-average risk, likely associated with summer vacations and increased road usage, which often result in longer travel distances and higher traffic volume. These findings highlight the importance of recognizing seasonal trends and planning accordingly to reduce accident rates.

The model's reliance on precipitation as the dominant predictor underscores the critical role of weather in road safety. Rain, snow, and icy conditions not only reduce road traction but also impair visibility and braking efficiency, contributing to an increased likelihood of accidents. Other influential factors, such as road-specific characteristics (e.g., high traffic areas) and temporal trends (e.g., holiday weekends), contribute to the overall risk assessment. These findings align with common expectations but provide actionable, data-driven guidance for drivers and policymakers alike. For drivers, the model suggests practical measures such as avoiding travel during extreme weather, reducing speed in wet or snowy conditions, and ensuring vehicles are equipped with proper winter tires and brakes during colder months. As for holiday seasons, drivers should be extra cautious due to the more congestion on the roads.

For policymakers and transportation planners, these insights are particularly valuable for improving resource allocation, expressing safety warnings to the public, and implementing

targeted inventions. For example, additional snow plowing and salting roads could be prioritized during the winter months, particularly in regions with higher accidents. Transportation agencies could also increase public safety messaging during holidays, encouraging drivers to plan their travel during off-peak hours and avoid driving in hazardous conditions. Furthermore, emergency response services could be better positioned in known high-risk areas during these peak periods, ensuring quicker response times and potentially reducing the severity of accidents.

Despite its strengths, the study has limitations. The focus on the 150 most accident-prone roads allowed the model to generalize well to high-traffic areas but limited its applicability to rural or less-trafficked roads, where accidents may still occur frequently but are underrepresented in the analysis. Additionally, the lack of traffic density data limited the ability to normalize accident counts fully and incorporating traffic density data would enable normalization of accident counts, providing more accurate risk assessment across both major and minor roads. Expanding the model to include real-time dynamic factors, such as live weather conditions, road closures, and traffic congestion levels, could significantly improve its predicting power. Integrating datasets on traffic volume and developing more complex models to capture interactions between variables are promising directions for further work.

Another area for exploration is the impact of real-time factors, such as current weather conditions or temporary road closures, on accident risk. Extending the analysis to include these dynamic elements could make the model more useful for predictive applications like route planning. At the same time, incorporation of socio-economic and behavioral data, such as driver demographics, vehicle types, and accident causes, to provide a more holistic analysis of accident risks could help in understanding how factors like distracted driving, fatigue, or vehicle age influence accident frequencies could allow policymakers to design targeted interventions, such as

public safety campaigns addressing driver behavior or incentives for regular vehicle maintenance. While it could be difficult to measure distracted driving, for example, it could be a great metric to have to further improve the model.

Future research could also investigate the spatial distribution of accidents to identify regional hotspots that require targeted interventions. For instance, analyzing accidents in urban centers like Northern Virginia and Richmond versus rural roads in the Shenandoah Valley could reveal critical differences in risk factors, such as traffic density, road quality, or driver behavior. This type of localized analysis would help policymakers design tailored safety measures for specific regions.

In conclusion, this study highlights the potential of machine learning techniques in analyzing road safety patterns and provides a strong foundation for both immediate practical applications and future research. By identifying high-risk months and roads, it equips drivers with the knowledge needed to make informed travel decisions while enabling policymakers to implement targeted strategies for reducing accident rates. The insights gained from this study can directly influence road maintenance schedules, public awareness campaigns, and resource allocation, ultimately contributing to safer travel conditions across Virginia. With further refinements, such as incorporating real-time data, behavioral metrics, and localized analysis, the findings have the potential to significantly reduce accident rates, improve driver safety, and enhance the overall effectiveness of transportation systems statewide. Through continued research and collaboration, data-driven approaches like this can play a crucial role in creating safer roads for all travelers.

References:

Machine Learning Results

National Highway Traffic Safety Administration. "Traffic Crashes Cost America Billions in 2019." NHTSA, 10 Jan. 2023, https://www.nhtsa.gov/press-releases/traffic-crashes-cost-america-billions-2019. .

Pre Analysis Plan - Machine Learning Final Project