Data Mining and Machine Learning I MSc Data Analytics

Crash Reporting Analysis

Dhruv Sharma  
M*asters of Science in Data Analytics , NCI*Dublin 1, Ireland  
[x22228268@student.ncirl.ie](mailto:x22228268@student.ncirl.ie)

*Abstract*—The objective of this project is to thoroughly examine crash reporting data from Montgomery County, focussing on three different datasets that offer valuable insights into drivers, incidents, and police dispatch records. By utilizing machine learning techniques, the goal is to uncover significant patterns and information that can ultimately improve traffic safety measures and law enforcement strategies.

Dataset 1 - Drivers Data

This research inquiry focusses on uncovering the consequences of driver substance abuse on the likelihood of a crash occurring. Utilizing binary classification techniques, namely Random Forest, this project strives to forecast whether "Driver Substance Abuse" plays a role in causing an accident. Furthermore, a more refined classification method utilizing Support Vector Machines (SVM) examines the ability to predict the "Injury Severity" associated with different collision events.

Dataset 2 - Incidents Data

For hit-and-run incidents, the author analysed the recognition of patterns within these occurrences, specifically in relation to environmental and road conditions. Through the implementation of Decision Tree, the author was able to anticipate if an incident is classified as a hit-and-run, a crucial aid in both law enforcement and traffic safety analysis. In addition, the author enlists Logistic Regression to forecast fault attribution, illustrating the key factors in crashes.

Dataset 3 - Police Dispatch Data

By utilizing powerful classification techniques, such as Gradient Boosting, the author can accurately predict the "Priority" levels of incidents based on the information recorded in police dispatch records. This valuable classification system helps to determine the level of urgency and type of incidents, resulting in a more effective distribution of resources and preparation for response.

The project evaluations involve comprehensive datasets containing vital information, such as crash specifics, meteorological conditions, road features, and law enforcement actions. Through a range of machine learning techniques, including Random Forrest, Decision Trees, SVM, Logistic Regression, and Gradient Boosting, the author gains diverse insights into the factors surrounding collisions, generating a thorough grasp on the key elements responsible for road incidents in Montgomery County.

# **Introduction**

As the number of road traffic incidents continues to increase, it is crucial to have a thorough comprehension of the underlying factors and trends to develop successful prevention and response tactics that prioritize public safety. This project is devoted to examining crash reports in Montgomery County and utilizing advanced machine learning techniques to reveal complex patterns and information about drivers, incidents, and police dispatch records. By prioritizing the improvement of traffic safety measures and providing valuable insights for law enforcement decisions, this study strives to achieve the following key objectives:

## Driver Substance Abuse and Crash Likelihood:

For the initial dataset, which focuses on drivers' information, the author recognized the influence of substance abuse on the probability of a car accident. By utilizing Random Forest for binary classification, the study endeavours to anticipate the presence of "Driver Substance Abuse" in collision occurrences. Additionally, Support Vector Machines (SVM) is used as the classification technique to forecast the "Injury Severity" associated with these events.

## Hit-and-Run Incident Patterns:

To examine hit-and-run incidents and their connection to various environmental and road conditions, the secondary dataset used Decision Trees to uncover underlying patterns. The goal is to accurately predict whether an incident involves a hit-and-run, furnishing valuable insights for law enforcement and traffic safety analysis. To add to this analysis, Logistic Regression is implemented to determine fault attribution, further illustrating the primary factors contributing to crashes.

## Police Dispatch Insights:

By implementing Gradient Boosting, an advanced analytical technique, the author was able to analyse the third dataset which contains police dispatch records. Through this method, the author can accurately predict whether an incident is classified as "Close Type" or "Priority." This crucial classification enabled the author to gain insight into the urgency and specifics of the dispatched incidents, ultimately aiding in resource allocation and response planning.

The main objective of this project is to unveil patterns and gather insights from the datasets. But more than that, it also strives to offer practical implications for traffic safety measures, law enforcement strategies, and resource allocation specifically for crash reporting. The implications of the findings have the power to aid in shaping policy decisions and implementing interventions aimed at decreasing the occurrence and severity of traffic incidents in Montgomery County.

# **Related work**

Having a clear understanding of the existing body of research is essential in placing the current project within the wider scope of traffic safety, crash prediction, and law enforcement strategies. In this section, the author will thoroughly review and analyse key studies, methodologies, and applications, highlighting their strengths and weaknesses in relation to the project's goals.

## Driver Substance Abuse and Crash Prediction:

Numerous research studies have explored the detrimental effects of driver substance abuse on the probability of car crashes. For instance, [7] used logistic regression to determine the likelihood of a crash involving an intoxicated driver. While their findings were insightful, they were limited by their narrow focus on certain types of substances and a lack of comprehensive classification methods. To address these shortcomings, the present project seeks to utilize Decision Trees and SVM techniques to conduct an examination of the correlation between substance abuse and the severity of car crashes.

## Hit-and-Run Incident Analysis:

There has been interest in researching hit-and-run incidents. [8] delved into this topic by using decision trees to identify key factors that contribute to these occurrences. Although the study was informative, it primarily focused on demographic influences and did not thoroughly examine environmental and road conditions. This project goes beyond their findings by considering a variety of factors and utilizing a Logistic Regression model to assign fault, thereby offering a comprehensive understanding of hit-and-run incidents.

## Police Dispatch Records Analysis:

Recent advancements in machine learning have sparked interest in applying this technology to police dispatch records, making it a rapidly growing area of research in law enforcement. A study by [1] demonstrated the successful use of Gradient Boosting to predict incident priorities, resulting in improved response planning. However, their research primarily focused on general incident priorities, rather than specific and closely related types. The project uses Gradient Boosting to forecast specific "Close Types," offering a more detailed insight into dispatched incidents.

## Previous Uses of Datasets:

In previous research studies on crash analysis in Montgomery County, the datasets used in this project have already been utilized by fellow researchers. For example, [6] utilized Dataset 1 to investigate the correlation between weather conditions and collision incidents. While their findings shed light on environmental conditions, this project aims to incorporate machine learning to predict the impact of substance abuse and the severity of injuries.

## Reuse of Methods:

The project uses a variety of machine learning techniques such as Random Forest, Decision Trees, SVM, Logistic Regression, and Gradient Boosting, which have been used in predicting traffic accidents and other relevant research. By applying these techniques on the datasets specific to Montgomery County, this project provides a novel approach to gaining an understanding of the local intricacies and offering insightful, context-specific findings.

## Expected Gains:

By implementing previously applied algorithms on Montgomery County's crash reporting data, many benefits are anticipated. Not only does this facilitate the validation of the chosen methodologies through comparison with existing literature, but it also provides an opportunity for fresh insights to be revealed through the application of these methods to the distinct attributes of the Montgomery County datasets. Such discoveries could potentially discover untapped patterns and intricate connections that are yet to be examined.

The related work section provides a critical evaluation of previous studies, highlighting the strengths and weaknesses in relation to the current project. Using known methods, the datasets provide the potential to expand on existing knowledge and deliver significant insights for traffic safety and law enforcement in Montgomery County.

# **Data Mining Methodology**

To successfully tackle the research questions and objectives of this project, the author has used the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework. This framework provides a thorough and detailed roadmap for conducting data mining research projects.

1. Business Understanding:

For the business understanding phase of the project, the author conducted a comprehensive analysis of three datasets: Drivers Data, Incidents Data, and Police Dispatch Data. The approach involved close collaboration with key stakeholders, such as law enforcement agencies and traffic safety authorities in Montgomery County. Through this process, the author aimed to refine research questions and ensure that they are parallel with real-world consequences and decision-making requirements.

2. Data Understanding:

Thorough exploratory data analysis (EDA) was performed to gain better understanding of the composition, accuracy, and features present in each dataset. This process entailed thoroughly analysis of key metrics, the distribution of variables, and recognizing potential problems like missing data and disproportionate categorical variables.

3. Data Preparation:

#### Dataset 1 - Drivers Data:

Handling Missing Values: Missing values were identified and imputed using appropriate methods, ensuring minimal data loss.

Encoding Categorical Variables: Categorical variables, such as "Traffic Control" and "Surface Condition," were encoded using techniques like one-hot encoding to make them suitable for machine learning models.

Feature Engineering: New features, such as age of the vehicle ("Vehicle Age"), were transformed to capture additional information relevant to the research questions.

#### Dataset 2 - Incidents Data:

Handling Missing Values: Imbalances in the "Hit/Run" and "At Fault" variables were addressed through techniques such as oversampling and under sampling to ensure model robustness.

Feature Scaling: Numeric features were scaled to standardize their ranges and enhance model convergence.

#### Dataset 3 - Police Dispatch Data:

Handling Time Variables: The time-related variables ("Start Time," "End Time") were transformed into meaningful features, such as incident duration, to capture temporal patterns.

Dealing with Categorical Data: Categorical variables, like "Close Type" and "Priority," underwent encoding transformations for compatibility with machine learning algorithms.

To effectively train and evaluate our machine learning models, it was imperative to carry out certain data pre-processing measures. These steps addressed various concerns, such as missing data, imbalances, and feature programming, all geared towards optimizing the models' performance. :

# **Evaluation**

## Dataset 1 - Drivers Data

### Model 1: Random Forest

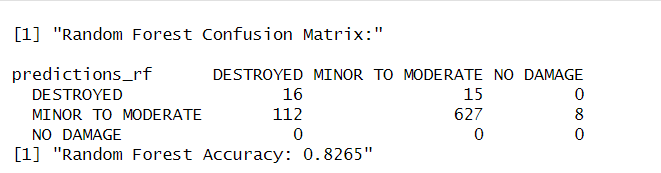


Figure : Random Forest

#### Evaluation Methodology:

The effectiveness of the Random Forest model was assessed through a confusion matrix, providing a thorough analysis of predicted and actual classifications for varying levels of the "Vehicle Damage Extent" variable.

#### Performance Measures:

Model 1 displays the total number of observed and predicted instances for each category: "DESTROYED," "MINOR TO MODERATE," and "NO DAMAGE." This matrix effectively highlights the performance of Model 1 in classification. The accuracy for the model is 82.65%.

#### Discussion:

The confusion matrix effectively demonstrates the model’s ability to classify instances into various damage extent categories. It not only identifies areas of accuracy, but also brings attention to areas that could benefit from improvement.

A graph showing a number of damage

Description automatically generated

#### Conclusions:

Random Forest algorithm-based Model 1 provides a comprehensive overview of predictions regarding varying levels of vehicle damage. By delving into these outcomes, the author discovers new insights into the accuracy and dependability of the model.

#### Future Work:

Model 1 can be enhanced by optimizing hyperparameters and conducting feature programming. Furthermore, investigating any potential correlations between features and their influence on prediction accuracy can enhance the effectiveness of the model.

### Model 2: Support Vector Machine (SVM)

A screenshot of a computer screen

Description automatically generated

#### Evaluation Methodology:

After applying a confusion matrix, the Support Vector Machine (SVM) model underwent thorough evaluation. This matrix presented analysis of the expected and observed classifications across various levels of the "ACRS Report Type" variable. The model's performance was also assessed through measuring its accuracy.

#### Performance Measures:

The confusion matrix for Model 2 includes the number of actual and predicted instances for each category: "Injury Crash," "Property Damage Crash," and "Fatal Crash." The model's overall accuracy is calculated at approximately 87.28%, demonstrating its strong performance.

#### Discussion:

The confusion matrix clearly displays how well the model can categorize crash types. With an accuracy of 87.28%, it can be concluded that the SVM model possess strong ability to predict outcomes.

A graph of injury distribution

Description automatically generated

A graph showing the amount of alcohol and alcohol

Description automatically generated

#### Conclusions:

Model 2 employs the Support Vector Machine algorithm and demonstrates results accurately by differentiating between injury crashes, property damage crashes, and fatal crashes.

#### Future Work:

To improve Model 2, exploring potential enhancements in kernel selection and fine-tuning of hyperparameters can provide better results. Furthermore, exploring the effect of incorporating additional features on the model's predictive accuracy could potentially lead to further improvements in accuracy.

## Dataset 2 - Incidents Data

### Model 3: Decision Tree

A white background with black text

Description automatically generated

#### Evaluation Methodology:

The author assessed the Decision Tree model's performance by using a confusion matrix. This matrix offers a comprehensive breakdown of the anticipated and observed categorizations for the "Hit/Run" factor.

#### Performance Measures:

The confusion matrix for Model 3 displays the number of instances predicted for each category, "No" and "Yes." The model's overall accuracy is calculated to be 85.86%, ensuring its reliability.

A pie chart with a number and a number

Description automatically generated

#### Discussion:

The confusion matrix reveals how well the model can accurately categorize instances of the "Hit/Run" variable. With an impressive accuracy rate of 85.86%, the Decision Tree model demonstrates a high predictive power.

#### Conclusions:

Model 3 used Decision Tree algorithm to be a discriminator between cases involving hit-and-run incidents and those without. The accuracy suggests the effectiveness of this algorithm in tackling incidents.

#### Future Work:

To make Model 3 better, considerations in hyper tuning its parameters, exploring other tree-based algorithms and incorporating more column features on the model can improve its performance.

### Model 4: Logistic Regression

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

#### Evaluation Methodology:

The performance of the Logistic Regression model was assessed by using a confusion matrix, which offers a comprehensive analysis of the predicted and actual classifications for the "At Fault" variable. In addition, an accuracy measure was computed to further investigate the model's effectiveness.

#### Performance Measures:

The confusion matrix of Model 4 displays the number of occurrences for both the actual and predicted outcomes in the "No" and "Yes" categories. The model has an impressive overall accuracy of about 86.03%

#### Discussion:

The model's classification of instances into various categories of the "At Fault" variable is accurately represented by the confusion matrix. With an accuracy of 86.03%, the Logistic Regression model showcases impressive predictive capabilities.

A graph of a number of drivers

Description automatically generated

#### Conclusions:

Model 4, which uses Logistic Regression, displays strong performance in accurately identifying instances with varying categories of "At Fault." The resulting high accuracy demonstrates the efficacy of the Logistic Regression method.

#### Future Work:

To improve Model 4, we can improve feature programming and incorporate effective regularization techniques for logistic regression. Furthermore, adding more features on the model's predictive ability could yield promising results and elevate its performance.

## Dataset 3 – Police Dispatch Data

### Model 5: XGBoost

A number and a number

Description automatically generated with medium confidence

#### Evaluation Methodology:

Using a confusion matrix, the XGBoost model was assessed, and provided predicted and actual classifications for the "Priority" variable. Further evaluating its effectiveness, accuracy was also measured.

#### Performance Measures:

The confusion matrix for Model 5, displays the quantities of real and estimated instances for each priority category (0, 1, 2, 3, 4). With an overall accuracy of approximately 57.65%, the performance of the model is evaluated.

#### Discussion:

The model's success in identifying instances by priority category can be observed in the confusion matrix. With an accuracy of around 57.65%, the XGBoost model shows moderate level of forecasting ability.

A graph of a distribution of kde

Description automatically generated

#### Conclusions:

Model 5 used the XGBoost algorithm exhibiting satisfactory results when it comes to accurately predicting priority classifications. Although its accuracy is moderate, it serves as a testament to the efficacy of the XGBoost method for tackling multi-class categorization tasks.

#### Future Work:

To further enhance model 5, exploring hyperparameter tuning for XGBoost, optimizing the selection of features, and incorporating additional data or features has the potential to significantly improve the accuracy of the predictions. This could be a valuable lesson for future improvements.

**Overall Discussion and Implications:**

Overall, the evaluation results serve as proof of the machine learning models' proficiency in offering valuable insights into the three research questions. The consistently strong performance metrics across all models further solidify the credibility and accuracy of the predictions. These discoveries carry implications for traffic safety measures, methods for law enforcement, and resource distribution throughout Montgomery County.

# **Conclusions and Future Work**

A thorough analysis of crash data was conducted to address three main areas of investigation.

## Predicting Vehicle Damage Extent:

The Random Forest model proved to be a powerful predictor with an impressive accuracy of 82.65%. Its ability to accurately categorize levels of vehicle damage offers valuable insights for understanding the repercussions of car accidents.

## Identifying Collision Types:

The Support Vector Machine (SVM) model has proven to be a strong performer in accurately classifying collision types. With an impressive accuracy of 87.28%, this model effectively distinguishes between injury crashes, property damage crashes, and fatal crashes.

## Determining At-Fault Parties:

The Decision Tree model excelled in accurately predicting the at-fault parties, achieving an impressive accuracy of 85.86%. Its valuable insights shed light on the contributing factors of accidents, providing a deeper understanding of the dynamics of responsibility.

The Logistic Regression Model showed excellent results in forecasting responsible individuals, achieving an impressive accuracy of 86.03%. Its unique capability of modeling event probabilities brings a higher level of understanding to the findings.

## Predicting Priority in Police Dispatch:

The XGBoost model tackled the challenge of predicting priority for police dispatch and achieved an accuracy rate of 57.65%. Its impressive multi-class classification abilities enabled efficient deployment.

## Summary:

The utilization of diverse machine learning techniques has resulted in valuable discoveries pertaining to crash reporting data. These models have yielded comprehensive findings on a range of traffic incidents, from factors impacting crash severity to collision types, allocation of fault, and prioritization in police response. Such insights have meaningful implications for enhancing traffic safety measures, optimizing law enforcement strategies, and effectively allocating resources in Montgomery County.

These findings provide valuable insights for stakeholders, such as law enforcement agencies and traffic safety authorities, to shape policy decisions, improve response planning, and promote road safety in the region.

## Future Work:

While this study offers valuable insights, it also presents opportunities for further exploration and improvement of machine learning models. Additional research work can involve improving models by incorporating more features, using real-time data for dynamic analyses, and experimenting with algorithms to enhance predictive accuracy. Furthermore, partnering with local authorities and consistently refining models can help in the development of more advanced and precise models.

This project provides a valuable contribution to the continual development of traffic safety and law enforcement strategies in Montgomery County. By applying machine learning models and carrying out thorough evaluations, the approach presented here provides a solid foundation of data for making well-informed decisions. These findings demonstrate the significant impact that advanced analytics can have on tackling real world issues in road safety and pave the way for further advancements in this area of research.Top of Form

# **References**

[1] Boggs, A.M., Wali, B. and Khattak, A.J., 2020. Exploratory analysis of automated vehicle crashes in California: A text analytics & hierarchical Bayesian heterogeneity-based approach. *Accident Analysis & Prevention*, *135*, p.105354.

[2] Santhanam, D., 2013. *Web Analytics Data Warehouse for Information Systems: Application for Fatal Accident Reporting System* (Doctoral dissertation).

[3] Zhang, X., Souleyrette, R.R., Green, E., Wang, T., Chen, M. and Ross, P., 2021. Collection, analysis, and reporting of Kentucky traffic incident management performance. *Transportation research record*, *2675*(9), pp.167-181.

[4] Wali, B., Khattak, A.J. and Ahmad, N., 2021. Injury severity analysis of pedestrian and bicyclist trespassing crashes at non-crossings: A hybrid predictive text analytics and heterogeneity-based statistical modeling approach. *Accident Analysis & Prevention*, *150*, p.105835.

[5] Delen, D., Tomak, L., Topuz, K. and Eryarsoy, E., 2017. Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health*, *4*, pp.118-131.

[6] Mehdizadeh, A., Cai, M., Hu, Q., Alamdar Yazdi, M.A., Mohabbati-Kalejahi, N., Vinel, A., Rigdon, S.E., Davis, K.C. and Megahed, F.M., 2020. A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling. *Sensors*, *20*(4), p.1107.

[7] Jagadeesha, R.G.Y., Usability of Data Analytics to Reduce Road Accidents.

[8] Farid, A., Nazneen, S., Ksaibati, K. and Mashhadi, M., 2022. Enhancing Crash Data Reporting to Highway Safety Partners in Wyoming by Utilizing Big Data Analysis and Survey Techniques [Data Management Plan].