Student name: Yingying Jiang

Student number: 501212156

Supervisor’s name: Ceni Babaoglu

Date of submission: June 4, 2023

Github link: https://github.com/catherine12jiang/CIND820\_capstone

Abstract

The problem is to develop a predictive model with the concentration on the injury severity of drivers in traffic collisions occurring on county and local roadways within Montgomery County. As for the theme, the researcher decides to choose classification and regression learning methods.

To be more specific regarding research problems, the researcher plans out 4 main research questions based on the summary of research questions listed above. Initially, what factors contributed to the injury severity in traffic collisions in Montgomery County? Secondly, how could a predictive model be effectively developed in order to forecast the injury severity level of drivers involved in traffic collisions? Furthermore, how do different modeling techniques and data preprocessing methods affect the accuracy of the predictive model? Lastly, how could the local policy makers, law enforcers, and the other potential stakeholders from Montgomery, Maryland utilize this model for enhancing road safety and reduce the injury severity of the drivers from traffic accidents?

The data used by the researcher is named as “Crash Reporting - Drivers Data” published by Montgomery County Government of Maryland, which can be sourced from the U.S. Government’s Open Data (data.gov). According to data.gov, this dataset provides detailed information on drivers involved in traffic collisions in Montgomery County, and the data was collected via the Automated Crash Reporting System (ACRS) of the Maryland State Police. (Data.gov, 2023) The raw data contains 159,357 rows and 41 columns. The data type of most columns can be converted to categorical data type, such as “Route Type”, “Collision Type”, “Weather”, “Surface Condition”, and “Light”.

The researcher proposes the following methodologies and techniques to tackle down each question listed from the second paragraph of this abstract. For the tech stack of this research project, Python and Jupiter Notebook are chosen considering the rising popularity among industrial practitioners specially in the data science field. Before conducting the research, the researcher suggests that there is a pressing need for data cleaning during data preparation considering the data is not in the well-maintained condition because of the inconsistencies occurred in several columns. Therefore, the researcher is implementing the following techniques in the data preprocessing stage: removing data columns with too many missing values, removing attributes with low variance, reducing highly correlated columns, balancing class attribute, feature selection, and processing PCA - Principal component analysis for Dimensionality Reduction. As for predictive modeling, Logistic Regression and Naive Bayes classifier are selected for investigating the first research question proposed by the researcher. As for the second and third research questions, Classification Tree and k-NN (k-Nearest Neighbours) are chosen by the researcher.

Literature review

Over years, there are numerous studies have been conducted based on vast variety of methodologies to investigate the relationship between traffic incident severity and its contributing factors. During the literature review stage, the author has read 19 journal articles, and the findings are as follows. 1. The researchers generally tend to define several categories for the contributing factors and group the factors under the defined categories. In this way, the data exploration is illustrated in an organized format. The common categories are driver, environment, road, vehicle, etc. 2. For the methodologies and tools used in those 19 studies, the author would like to address the modeling algorithms used in the studies. Firstly, modeling algorithms that commonly used in the investigation of the traffic accidents contributing factors are K-NN, Naïve Baye, logistic regression, and Decision Tree. However, several researchers from the Asia and South America used the Support Vector Machine (SVM) and a relatively new approach called Gradient Boosting Machines, in which the author has limited knowledge. During the next stage, the author will attempt to utilize these two relatively less popular methods and review the performances accordingly. At the same time, the author found out that multiple journals demonstrated quite robust methodology and strong statistical knowledge, but the conclusion from them are perfectly aligned with ordinary people’s common sense. For example, one of the journal studies used SVM and other modeling tactics to build a machine learning model. For their conclusions and recommendations, they suggested the driver to be more careful when driving with limited light condition, which seems pointless for conducting a multi-stage research study because most ordinary people already know very well about this fact.

In following paragraphs, the author selected 6 journals which are either strongly related to the research questions of this study or demonstrated an extremely detailed implementation tactics during each step of implementation. Moreover, a relatively summary was also generated for each chosen article.

In the study carried out by the Kongju National University in South Korean, the researchers aimed to investigate the relationships between major factors and types affecting the injury severity in the traffic accidents in Seoul city. This research is particularly significant in the literature review section considering it has a quite similar research objective with the author’s study. According to Jeong, Kim, and Han, the study derived several critical factors that were found to affect the severity of traffic accidents via supervised learning methods (i.e., ensemble-based and regression-based algorithms) and discovered dominant accident types via unsupervised learning methods. (Jeong et al., 2022). According to the research team's findings, the conclusion contrasts with conventional mindset as it reveals that factors such as season, day of the week, and even weather have been determined to lack a significant influence on the severity of traffic accidents in Seoul city. The value of the journal written by the South Korean research team extends beyond the methodologies and implementations. In particular, the significance lies in the fact that the conclusion drawn from the machine learning model contradicts the common beliefs or intuitive expectation, which encourage the other researchers to explore the other unconventional factors involved in this research field.

There was another study regarding machine learning models related to road traffic accidents conducted by the researchers from University of Johannesburg. According to Bokaba, Doorsama, and Paul, the study aimed to assess prediction model designs for RTAs (road traffic accidents) to assist the local transport authorities and policy makers. (Bokaba et al., 2022) Indeed, the study utilized the dataset collected from Gauteng Province in South Africa and employed several classic supervised machine learning methods. Although the study conducted in South African is not directly relevant to the injury severity, the research team extensively demonstrated the use of various tools such as K-NN, naïve Bayes, logistic regression, etc. It is also valuable that the research team dedicate numerous pages on describing the dimension reduction and machine learning model performance evaluation. The methodologies are commonly used in every step of the data science project life cycle and could potentially be adapted or applied in the author’s own study.

The study carried out by the collaboration between the scholars from China and United States of America has a concentration of investigating the factors affecting passenger car and truck drivers’ injury severity in the traffic collisions, which has a strong connection related to the author’s study. Moreover, the research team adopted a method to reset the levels for the categorical variables in order to avoid an excessively large number of levels. As a result, each attribute has meaningful and manageable categories, which sets a solid step for the further development of the machine learning model. Meanwhile, the class variable – injury severity is highly imbalanced in the dataset where the no injury is significant dominant than others. To cope with imbalanced data, the researchers used four unique methods to conduct the proper treatments in order to enhance the performance of classifier, which inspires the author to a great extend because the dataset from Montgomery, Maryland is significantly imbalanced as well. However, the findings from this journal were limited in their significance due to their alignment with common knowledge. Despite the differing perspectives regarding the study’s findings, the author found value and relevance in the methodologies and approaches used by the research team.

The researchers from Chang'an University conducted a study that focused on driving experiences and other potential factors contributing to traffic collision severity. Unlike other journals in the literature review section, the research team utilized various modeling and data cleaning techniques that are not commonly mentioned by industrial practitioners. In particular, they employed techniques such as SMOTE (Synthetic Minority Over-Sampling Technique), ENN (Edited Nearest Neighbor), and Tomeklink for data resampling. This was necessary because the dataset they used had imbalanced injury severity categories. By applying these techniques, they aimed to address the data imbalance issue and improve the reliability of their results. Additionally, the researchers dedicated significant efforts to the boosting algorithm in order to further improve the performance of their model. This approach inspired the author to include a section on boosting algorithms in their future study. Based on the result of the machine learning process, the researching team illustrated that the pavement surface conditions, overload condition, and gender have the least impact on the accident severity, which contradicts from ordinary people’s common perceptions.

One of the journals that related to the author’s study is carried out by a researching team from United Kingdom. The focus of this study is laid on building a machine learning model to analyze the accident severity’s contributing factors. The UK team obtained a raw data with 60 columns and 135453 rows, and after data preprocessing, there were only 16 attributes left. The author was in the same dilemma during the data cleaning process, and this study from UK provided insightful opinions when the data scientists must decide which attributes needs to be trimmed off. Meanwhile, the study’s tech stack is 90% aligned with the author’s decision. Especially, there are several parts contain comprehensive information about using Scikit Learn and Pandas during the data modelling and evaluation stages. For the conclusion of the journal, the UK team pointed out that the vehicle factor is the most significant contributing factor to the traffic accident severity in UK.

Also, there is another journal focused on developing a machine learning model to investigate the car accident severity and casualties. What makes this study stand out is that the researchers did not simply remove the null value rows. Instead, the data was imputation by a using method named as Hot-deck imputation, which is quite beneficial to the author of this study. At the same time, another data cleaning method called interquartile method was applied to clean the numeric attributes from the dataset. During the implementation of the data modeling, the researchers used three different machine learning algorithms naïve baye, multinomial logistic regression, and random forest classifier to build a global model. This approach potentially enhanced the accuracy of the model. However, the data generated by naïve baye algorithm was taken off from the global model due to its low accuracy during the evaluation stage. As for the conclusion, apart from the points can align perfectly with common sense, the machine learning model is suggesting that Snowing with high winds weather is least likely to cause serious traffic accidents. It is quite exciting to review the journals that contain conclusions which are challenging ordinary people’s common perception.

Although there are numerous studies have been conducted regarding similar research questions as the author proposed, the research is still worth pursuing because of the following reasons.

1. Continuously evolving methodologies: The field of research methodologies, specifically in the machine learning field, is continuously changing and advancing. By comparing the journals from 2013 and 2023, there are numerous differences between the researcher’s implementations. By adapting to the new methodologies, there could be more findings uncovered by the researchers even when it comes to the same topic.
2. Enhancing of machine learning frameworks: The machine learning packages and frameworks are evolving at the same time. There are cases where certain functions are only supported only supported from specific version codes onwards. Therefore, the author’s study is worth for the whole data science community for serving as a reference about utilizing the latest version of a machine learning framework like scikit-learn.

In summary, although the study could potentially have certain parts overlapped with the previous research, conducting this study does hold the value and the relevance because of the continuously evolving methodologies and machine learning frameworks.

Data Cleaning

The original dataset contains 43 columns including 4 numeric columns and 39 object columns. The columns names from raw data are appropriate and meaningful, therefore, the author will keep original columns names.

During the data cleaning process, the author found out there are multiple columns consist over 90% of null values. Therefore, the columns were removed because of too many missing values. Also, there are many rows containing records such as unknown and others. Depending on the column name and the nature of the traffic collision, the rows were either removed from the dataset or imputed by using forward fill tactics. Moreover, several columns contain large number of unique values, and the author decided to convert them into a manageable size by grouping them with less levels. For time variable, it is a special case because the author decided to convert it into two new columns named “Day of Week” and “Time of Day” and dropped the original date column due to the excessive number of levels.

After the removal of unrelated columns, meaningless columns, and columns with too many missing values, all the columns are converted to categorical type since object type would consume more resources during the data modeling stage.

Data Description

After data cleaning, this is adjusted data frame’s dictionary. This data set totally consists of 105,379 rows and 14 columns. As for the target variable, the chosen attribute is “Injury Severity” based on the research questions from the abstract section. It is also noticeable that the class variable is significantly imbalanced due to the nature of traffic collisions.

|  |  |  |
| --- | --- | --- |
| Column Name | Description | Variable type |
| Route Type | Type of the Roadway at crash location. | Categorical |
| Collision Type | Type of traffic collision | Categorical |
| Weather | Type of the weather at collision location | Categorical |
| Surface Condition | Condition of roadway surface | Categorical |
| Light | Condition of lighting | Categorical |
| Traffic Control | Signage or traffic control devices | Categorical |
| Driver Substance Abuse | Substance abuse detected for drivers involved | Categorical |
| Driver At Fault | Whether the driver was at fault | Categorical |
| Injury Severity | Severity of injury to the driver | Categorical |
| Speed Limit | Local area posted speed limit | Categorical |
| Location | Exact location of the collision | Categorical |
| Day of Week | The day of week when collision happened | Categorical |
| Time of Day | The time of day when collision happened | Categorical |
| Driver Distracted By | Whether the driver was distracted | Categorical |

As for each column’s value distribution, the author plotted following frequency bar charts. The charts themselves are relatively self-descriptive.

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, font, parallel

Description automatically generated

A picture containing text, screenshot, font, diagram

Description automatically generated

A picture containing text, screenshot, diagram, number

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, font

Description automatically generated

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

A picture containing text, screenshot, diagram, display

Description automatically generated

A picture containing text, screenshot, font, parallel

Description automatically generated

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Additionally, the spearman correlation analysis was completed after the author generated a new data frame that was converted from the cleaned data set to numeric values on basis of the index of the levels. Because Spearman correlation coefficients does not require the data to be normally distributed and based on the ranks of the values instead of the actual values. By using seaborn and matplot, the heatmap of Spearman correlation analysis results are presented in a well-organized approach.

A screenshot of a graph

Description automatically generated with low confidence

The analysis of the provided table reveals two notable correlations. First of all, the positive correlation between "Route Type" and "Speed Limit" indicating that certain types of routes tend to have higher speed limits. This finding implies that the nature of the route may influence the designated speed limits in the specific area. Additionally, a positive and significant correlation between "Weather" and "Surface Condition" implying that specific weather conditions are associated with particular surface conditions. This correlation suggests that certain weather patterns, may impact the condition of the road surface, potentially affecting driving conditions and safety considerations.

Research Methodology

As for the research methodology, the author aims to employ an iterative data science development cycle to conduct the entire research. This methodology is inspired by the popular software development philosophy named Agile, which involves a flexible and iterative approach that allows adaptation, changes, and improvements throughout the entire research process.

By adopting this methodology, the author could break down the entire research into smaller manageable iterations. Each iteration involves data preprocessing, data modeling, and performance evaluations. After each iteration, the author can make the adjustments in order to enhance the performance of the modeling and ensure the research strictly aligned with the research questions and objectives.

To provide a more visualized approach for understanding the methodology of this research, the author created a flow chart that serves as a visual roadmap to guide the research process.

A diagram of a flowchart

Description automatically generated with low confidence

Reference:

Bokaba, Doorsamy, W., & Paul, B. S. (2022). Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents. *Applied Sciences*, *12*(2), 828–. https://doi.org/10.3390/app12020828

Chen, Shao, H., & Ji, X. (2021). Insights into Factors Affecting Traffic Accident Severity of Novice and Experienced Drivers: A Machine Learning Approach. International Journal of Environmental Research and Public Health, 18(23), 12725–. https://doi.org/10.3390/ijerph182312725

data.montgomerycountymd.gov. (2023, May 13). *Crash reporting - drivers data*. data.gov. https://catalog.data.gov/dataset/crash-reporting-drivers-data

Jeong, Kim, I., Han, K., & Kim, J. (2022). Comprehensive Analysis of Traffic Accidents in Seoul: Major Factors and Types Affecting Injury Severity. Applied Sciences, 12(4), 1790–. https://doi.org/10.3390/app12041790

Megnidio-Tchoukouegno, & Adedeji, J. A. (2023). Machine Learning for Road Traffic Accident Improvement and Environmental Resource Management in the Transportation Sector. Sustainability (Basel, Switzerland), 15(3), 2014–. https://doi.org/10.3390/su15032014

Pourroostaei Ardakani, S., Liang, X., Mengistu, K. T., So, R. S., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road car accident prediction using a machine-learning-enabled data analysis. *Sustainability*, *15*(7), 5939. https://doi.org/10.3390/su15075939

Wahab, & Jiang, H. (2019). A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PloS One, 14(4), e0214966–e0214966. <https://doi.org/10.1371/journal.pone.0214966>

Zhou, B., Wang, X., Zhang, S., Li, Z., Sun, S., Shu, K., & Sun, Q. (2020). Comparing factors affecting injury severity of passenger car and truck drivers. *IEEE Access*, *8*, 153849–153861. https://doi.org/10.1109/access.2020.3018183