TRAFFIC COLLISION AND VICTIM ANALYSIS: A CASE STUDY ON PREDICTIVE MODELING AND INSIGHTS

Exploring Machine Learning Applications for Traffic Safety and Victim Impact Assessment

SENG 550 PROJECT

Phuong Le (30175125) – Phuong.le1@ucalgary.ca
Sean Tan (30094560) - sean.tan3@ucalgary.ca
Victor Campos (30106934) - victor.campos@ucalgary.ca
Kyle Hasan(30087880) - kyle.hasan@ucalgary.ca

Table of Contents

Preamble

Contribution of Team Members:

- Sean Tan: Loading & preprocessing data, Logistic Regression modelling and visualizations for predicting INJURY_SEVERITY_BINARY
- Phuong Le: Linear regression modeling for predicting VICTIM_AGE
- Kyle Hasan: Loading and cleaning data, feature engineering and removing invalid data and making sure data was in a format where it could be used by ML models.
- Victor Campos: Cleaning data, Logistic regression modeling and visualizations for predicting VICTIM_EJECTED

Declaration

We, the undersigned, declare that the statement of contributions and estimate of total contribution is true.

Signed by: Phuong Le, Sean Tan, Victor Campos, Kyle Hasan

Repository

https://github.com/seantan88/SENG550-Final-Project.git

Abstract

This project analyzes the California Traffic Collision dataset to explore patterns and build predictive models. Our focus is on predicting critical attributes, including victim age, injury severity, and if the victim was ejected from the vehicle during the crash. Initial analysis involved data cleaning, handling missing values, and feature engineering. First, by using linear regression, we predicted VICTIM_AGE, achieving an R^2 score of 0.32222, highlighting the limitations of feature relevance and data variability. Logistic regressions were used to predict the severity of injuries sustained from a collision and whether the victim was ejected from the vehicle during the accident, which yielded AUC values of 0.785 and 0.94 respectively. Feature importance analysis revealed significant predictors such as victim seating position and safety equipment usage, providing actionable insights for future improvements. This work underscores the potential and challenges of predictive modeling in traffic safety analysis, offering a foundation for more comprehensive studies.

Introduction

Problem Selection

Traffic collisions result in significant societal and economic impacts. Identifying patterns and predicting attributes such as accident severity, victim demographics, and accident timing can inform preventative measures and improve road safety.

Dataset

The project uses two datasets from the California Traffic Collision dataset on Kaggle:

- **Collision Records:** Detailed records of collisions, including information such as `CASE_ID`, `CHP_BEAT_TYPE`, `PRIMARY_RD`, `DIRECTION`, `INTERSECTION`, and `COLLISION_TIME`.
- **Victim Records:** Detailed victim data, including `VICTIM_AGE`, `VICTIM_SEATING_POSITION`, `VICTIM_EJECTED`, `VICTIM_SEX`, `VICTIM_SAFETY_EQUIP1`, and injury details.

Both datasets underwent cleaning and preprocessing to handle missing values, standardize features, and merge data based on `CASE_ID`.

Importance of the Problem

Analyzing traffic collision data is crucial for reducing accidents and fatalities. Insights derived from predictive modeling can help policymakers and engineers design safer road systems.

Literature Review

Analyzing traffic collision data is a well-researched area in transportation safety and urban planning. Prior studies have utilized machine learning models such as Random Forests and Gradient Boosting Machines to predict accident injury severity based on variables such as road category [2]. Temporal analyses have highlighted different patterns based on variables such as day of the week [2][1] and hour [1], while geospatial data has been used to identify high-risk locations (hotspots) [1]. There have also been studies using Random Forest models to find areas where accidents are more likely to happen and less dangerous routes based on historical accident data [3].

Gaps Existing and Addressed

1. While prior studies provide valuable insights, specific gaps remain:

- **Feature-Level Understanding**

  Many studies emphasize accident severity and temporal patterns but do not delve deeply into feature-level relationships, such as how **VICTIM_AGE** correlates with factors like seating position or safety equipment usage. Addressing this gap is crucial for understanding victim-specific risks and tailoring safety interventions.

- **Victim-Specific Insights**

  Existing works often overlook individual victim characteristics. Predicting attributes like **VICTIM_AGE** using collision-related features remains underexplored in the literature.

- **Explainability in Predictive Models**

  Most machine learning models prioritize accuracy but lack interpretability. Understanding the importance of specific features, such as safety equipment or temporal patterns, can enhance the application of these models in policy making and safety improvements.

By focusing on these gaps, our project builds a **Linear Regression Model** to predict **VICTIM_AGE**, evaluate feature importance, and uncover temporal accident patterns. This approach contributes to the field by linking victim-specific factors to traffic collision dynamics.

2. Evaluating feature importance in linear regression models.
3. Investigating temporal accident patterns for safety insights.

Data Analysis Questions

1. What are the key features influencing victim age?
2. What factors influence the probability of a collision resulting in severe or fatal injuries, and how well can they predict the severity of injuries sustained?
3. What factors predict accident severity?

Proposal and Findings

This project proposes a combination of linear regression and classification models to analyze the California Traffic Collision dataset, focusing on understanding victim-specific attributes and collision outcomes. Key findings include:

1. Victim Age

   Moderate Success in Predicting Victim Age:

   Linear regression achieved an R^2 score of 0.32222, highlighting some predictive power despite the inherent challenges of feature relevance and data variability.

   Identification of Significant Features:

   Feature importance analysis revealed key predictors for VICTIM_AGE, such as:

   - Victim seating position (positive impact).
   - Safety equipment usage (notably VICTIM_SAFETY_EQUIP2 and VICTIM_EJECTED).
   - Gender of the victim (negative impact).

   These insights contribute to understanding victim demographics in traffic collisions.

2. Injury Severity

   Reasonable Predictive Power in Estimating Injury Severity:

   The first logistic regression achieved an AUC value of 0.785, indicating a good ability to predict severe injuries from the selected features.

   Identification of Significant Features:

   Key features in predicting INJURY_SEVERITY_BINARY included:

   - Ejection from vehicle (strong positive correlation with severe injury)
   - Safety equipment usage (strong negative correlation with severe injury)
   - Intersection (lower injury severity at intersections)

3. Victim Ejected

   Great Predictive Power in Estimating Victim Ejection:

The first logistic regression achieved an AUC value of 0.969, indicating a great ability to predict wether a victim was ejected using the selected features.

Key features in predicting VICTIM_EJECTED_BINARY included:

- Seating position within veichile had a strong relationship with the likelihood of ejected status.
- Safety equipment usage also had a strong relationship with the likelihood of ejected status
- CHP beat type had a less important and pronounced impact, but its presence improved model accuracy when combined with other features.

Methodology

Exploration of Data Features

1. **Data Cleaning**:
    - Handled missing values in both collision and victim datasets.
    - Dropped rows with null values.
    - Dropped rows with invalid values like negative distances.
    - Convert age 999 to age 0 since this is age 0 and remove removes with age 998 since this is age unknown.
    - Removed unused columns, such as REPORTING_DISTRICT.
    - Filter out rows with invalid values such as rows with PRIMARY_RD or SECONDARY_RD containing “...”.
    - Combined datasets using CASE_ID to provide a comprehensive view of collision and victim details.
2. **Feature Engineering**:

- Transformed categorical variables, such as VICTIM_SEATING_POSITION, VICTIM_SAFETY_EQUIP2, and PRIMARY_RD, into numerical features using StringIndexer and OneHotEncoder.
- Standardized numerical variables (e.g., COLLISION_TIME) with StandardScaler.
- Combined all features into a single feature vector using VectorAssembler.Feature Selection:

Experiment Setup

1. **Target Variables**:
    - VICTIM_AGE was selected as the target variable for regression to explore relationships between collision details and the age of victims.
        - **Purpose:** Understanding how collision factors correlate with the age of victims can help in tailoring safety measures, designing better safety policies, and understanding demographic-specific risks
    - INJURY_SEVERITY_BINARY was constructed from the VICTIM_DEGREE_OF_INJURY variable, and selected as the target variable in our first logistic regression.
        - **Purpose:** Understand how factors related to the victims and collision influence the chance of sustaining severe injury from an accident. This helps us isolate which factors have the largest impact on injury severity, allowing us to design policies around these findings to improve road safety.
    - VICTIM_EJECTED_BINARY was selected as a target variable for our second logistic regression
        - **Purpose:** Understand how factors related to victims and collisions influence the likelihood of being ejected during an accident. This helps identify the most impactful factors, enabling the design of better safety features and policies to prevent ejection and reduce injury risks.

Models Used:

- Linear regression was chosen to predict the continuous targets.
- Logistic regressions were used to predict binary targets.

2. **Data Split**:
    - Linear Regression: The dataset was split into 80% training and 20% testing subsets. A fixed seed was used to ensure reproducibility.
    - Logistic Regression: The dataset was split into 70% training and 30% testing subsets. A fixed seed was used again to ensure reproducibility.

Experimentation Factors

ML Algorithm: Linear Regression with L2 regularization

1. Hyperparameters: No additional hyperparameter tuning was performed beyond defaults
2. Evaluation metrics:

    Regression metrics were used to assess model performance:

    - R^2 (coefficient of determination)
    - Root Mean Squared Error (RMSE)
    - Mean Absolute Error (MAE)

ML Algorithm: Logistic Regressions

1. Hyperparameters: Tuning brought no improvements to model capability, and was thus was not performed.
2. Evaluation Metrics:
    - AUC (Area Under Curve)
    - Accuracy
    - Precision
    - Recall

Experiment Process

- Prepared the data pipeline, including indexing, encoding, scaling, and assembling features.
- Trained the Linear Regression model on its training dataset, trained the logistic regression models on their respective training sets.
- Evaluated the models on the testing dataset using the chosen metrics.
- Analyzed feature importance using the model coefficients.

- Conducted residual analysis to diagnose potential model weaknesses.

Performance Metrics

Linear Regression

- **R^2:** 0.32069
    - Indicates that ~32% of the variance in VICTIM_AGE is explained by the features.
- **RMSE:** 16.18
    - Reflects the average magnitude of error in predicted victim ages.
- **MAE:** 12.56
    - Represents the average absolute error between actual and predicted victim ages.

Logistic Regression (INJURY_SEVERITY_BINARY)

- **AUC:** 0.78545
    - Indicates that the model predicts injury severity correctly ~78.5% of the time
- **Precision (by label):** Non-severe → 0.81089, Severe → 0.76387
    - Indicates that the model correctly predicts non-severe injuries ~81% of the time, and severe injuries ~76% of the time
- **Recall (by label):** Non-severe → 0.74448, Severe → 0.82642
    - Indicates that the model captures ~74% of the true non-severe injures, and ~82% of the true severe injuries
- **Accuracy:** 0.78545
    - This tells us that the overall accuracy of the model lies around ~78%

Logistic Regression (VICTIM_EJECTED_BINARY)

- **AUC**: **0.96906**
    - Indicates that the model predicts injury severity correctly ~96..9% of the time
- **Precision (by label):** Non-ejected → 0.0.92624, Ejected → 0.0.92272
    - Indicates that the model correctly predicts 92.6% of non-ejected victims. The model accurately predicts 92.3% of ejected victims.
- **Recall (by label):** Non-ejected → 0.92237, Severe → 0.92657

    - The model captures 92.2% of actual non-ejected victims, and successfully identifies 92.7% of actual ejected victims.
- **Accuracy:** 0.96906
    - The overall accuracy of the model is approximately 96.9%, indicating strong performance across both classes.

Results

Key Findings in Exploratory Data Analysis and Prediction

1. **Feature Importance**:

   Linear Regression:

   - **VICTIM_SEATING_POSITION_vec** and **VICTIM_SAFETY_EQUIP2_vec** had the highest positive coefficients, indicating significant relationships with VICTIM_AGE.
   - V**ICTIM_EJECTED_vec** and **VICTIM_SEX_vec** showed notable negative influences.
   - Numerical feature **COLLISION_TIME** had a small but measurable negative impact.

   Logistic Regression (INJURY_SEVERITY_BINARY)

   - **VICTIM_EJECTED_vec** and **VICTIM_SEATING_POSITION_vec** had the largest positive coefficients.
   - **INTERSECTION_vec, VICTIM_SAFETY_EQUIP2_vec,** and **DAY_OF_WEEK_vec** had particularly strong negative coefficients.
   - **DIRECTION_vec** had a smaller and less noticeable impact on the severity of injuries sustained from an accident.

   Logistic Regression (VICTIM_EJECTED):

   - **VICTIM_SEATING_POSITION_vec** and **VICTIM_SAFETY_EQUIP1_vec** had the highest positive coefficients, indicating a strong relationship with the likelihood of **VICTIM_EJECTED_BINARY** (ejected status).

- **VICTIM_SAFETY_EQUIP2_vec** and **VICTIM_SEX_vec** showed smaller, yet notable, negative influences on the probability of ejection.
- **CHP_BEAT_TYPE_vec** had a less pronounced impact, but its presence improved model accuracy when combined with other features.

2. **Model Predictions**:

   - **The linear regression model** achieved an R^2 of 0.322, indicating moderate predictive power. This result suggests that the current feature set explains only about 32% of the variance in VICTIM_AGE. While it identifies some significant patterns, the model's ability to generalize is limited. Prediction errors highlighted limitations in capturing the full variance in victim age.
   - **The logistic regression model (INJURY_SEVERITY_BINARY)** showed reasonably accurate predictive power, as it yielded an AUC value of 0.785. This result signifies that the model is able to correctly identify and classify the severity of injury.
   - **The logistic regression model (VICTIM_EJECTED)** demonstrated strong performance, achieving an AUC score of 0.969 on the test data. This result indicates that the model has excellent discriminative power in predicting whether a victim was ejected or not during a collision

Model Diagnostics

Linear Regression

- Residual analysis showed no severe skewness but revealed heteroscedasticity, indicating variability in errors across the range of predictions. This suggests that the relationship between features and the target variable is not perfectly linear and that other unmodeled factors may be at play.
- Features like seating position and safety equipment were significant, but others may need refinement or inclusion to improve the model's predictive power.

Logistic Regression (INJURY_SEVERITY_BINARY)

- The target variable faced severe imbalancing issues, but this was rectified through the use of class weights
- While some features like VICTIM_SAFETY_EQUIP2_vec and VICTIM_SEATING_POSITION_vec showed significant impacts on injury severity, there are others (i.e DIRECTION) that could be refined or replaced to improve model performance.
- The creation of the target variable during the ML process seemed to create some null values, which created the need to drop null values an additional time during the ML process.

Logistic Regression (VICTIM_EJECTED)

- Precision remained consistently high (~98%) at lower recall values, indicating strong confidence in correctly identifying non-ejected victims.
- The area under the curve (AUC = 0.969) reinforces the model's robust predictive performance.
- A significant class imbalance was detected, with Class 1 (Ejected) comprising ~4% of the total dataset. Re-weighting was applied, resulting in a weight of 12.4 for minority class samples and 0.52 for the majority class. This adjustment mitigated imbalance issues and improved model accuracy.

Conclusions

Linear Regression

1. **Purpose of Predicting `VICTIM_AGE`:**
    - The analysis highlights how collision circumstances relate to victim age, providing insights into age-related vulnerabilities during accidents.
    - These insights can guide policymakers, emergency services, and safety researchers to develop targeted interventions for specific demographics.
2. **Strengths:**
    - Successfully identified key victim-specific factors influencing VICTIM_AGE.
    - Demonstrated the utility of linear regression for initial exploratory analysis.
3. **Limitations:**

- Moderate performance metrics suggest limitations in feature space and the linear model's assumptions.
- Variability in data and unmodeled interactions may affect accuracy.

4. **Future Directions:**
   - Incorporate additional features like weather conditions or time of day.
   - Experiment with non-linear models or ensemble methods to capture complex patterns.
   - Investigate how findings can directly translate into safety measures or policy enhancements.

Logistic Regression (INJURY_SEVERITY_BINARY)

1. **Purpose of Predicting `INJURY_SEVERITY_BINARY`**:
   - Illustrate the effects characteristics of victims or collisions have on the severity of injury sustained during an accident.
   - Better understand the risk factors associated with car accidents to make well informed policy and regulation decisions.
2. **Strengths:**
   - Achieved a high **AUC score of 0.785**, demonstrating reasonable model performance.
   - Successfully identified significant features, particularly seating position and victim ejection which both strongly influence the severity of injury.
3. **Limitations:**
   - If we had access to other data like locational data, weather data or more information about the size and speed of the cars involved in the crash, we could improve model performance
4. **Future Directions:**
   - Include interaction terms (i.e VICTIM_SAFETY_EQUIP1 x VICTIM_AGE) for a more robust set of features

Logistic Regression (VICTIM_EJECTED)

5. **Purpose of Predicting `VICTIM_EJECTED`**:
   - The logistic regression model highlights critical factors contributing to the likelihood of victim ejection in vehicle collisions.

- Understanding these relationships can guide road safety initiatives, inform first responders, and improve safety equipment design.

6. **Strengths:**
   - Achieved a high **AUC score of 0.969**, demonstrating excellent model performance.
   - Successfully identified significant features, particularly seating position and safety equipment, which directly influence victim ejection.
7. **Limitations:**
   - Limited feature set excludes important external factors like collision speed, weather conditions, or vehicle type, which could enhance predictive power.
8. **Future Directions:**
   - Incorporate additional features like weather conditions, vehicle-specific data, and collision severity to increase the sophistication of the model.

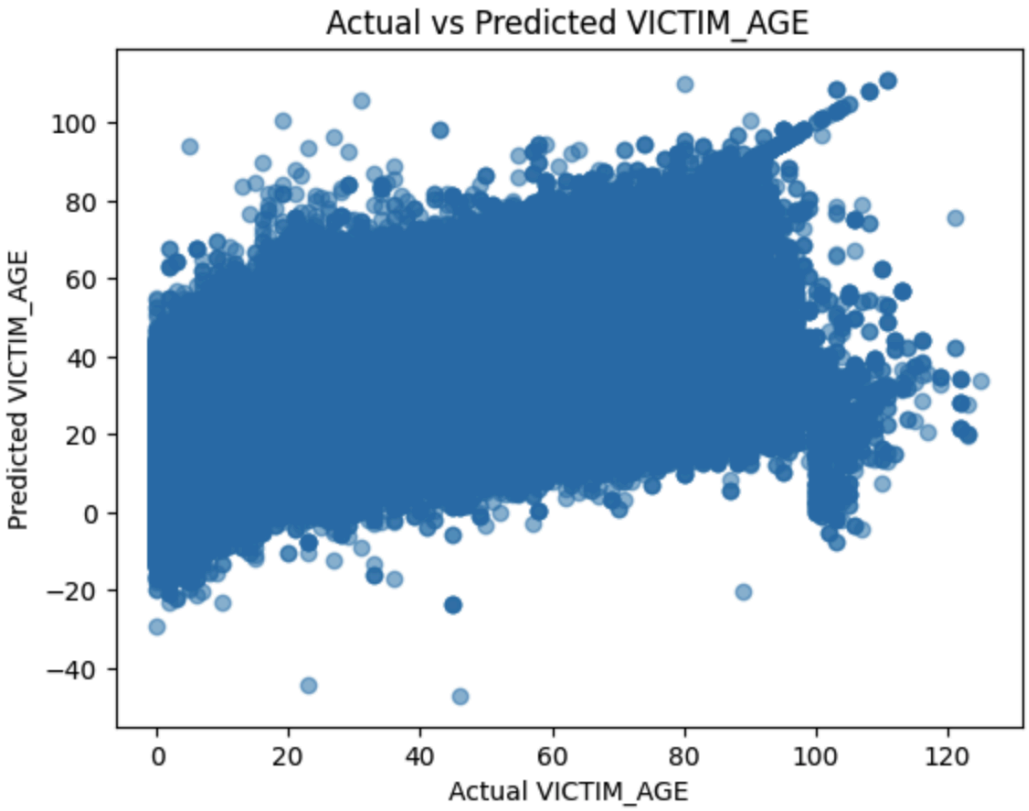
References

[1] Esri, “Analyzing traffic accidents in space and time,” desktop.arcgis.com. https://desktop.arcgis.com/en/analytics/case-studies/analyzing-crashes-1-overview.htm (accessed Dec. 16, 2024).

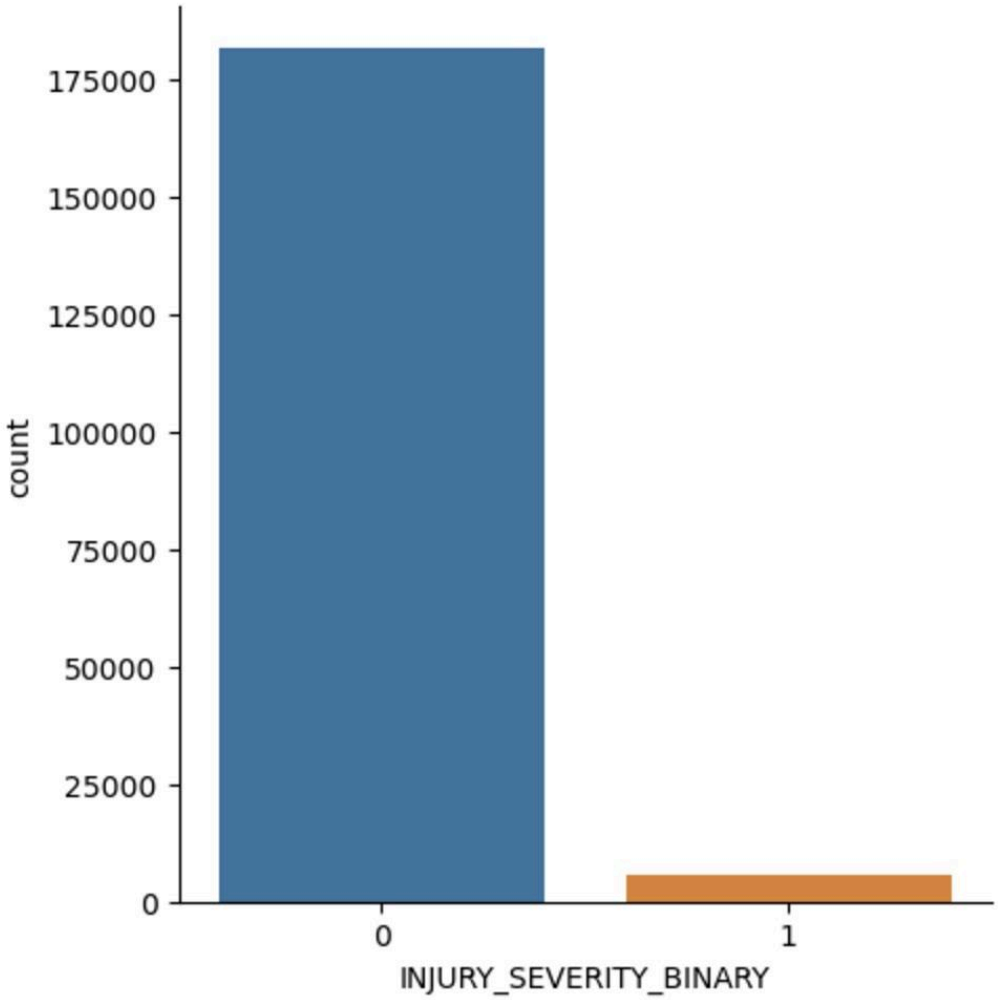
[2] S. Ahmed, M. A. Hossain, S. K. Ray, M. M. I. Bhuiyan, and S. R. Sabuj, “A study on road accident prediction and contributing factors using explainable machine learning models: Analysis and Performance,” sciencedirect.com. https://www.sciencedirect.com/science/article/pii/S2590198223000611 (accessed Dec. 16, 2024).

[3] Y. Berhanu, D. Schröder, B. T. Wodajo, and E. Alemayehu, “Machine learning for predictions of Road Traffic Accidents and spatial network analysis for safe routing on Accident and congestion-prone road networks,” sciencedirect.com. https://www.sciencedirect.com/science/article/pii/S2590123024009927 (accessed Dec. 16, 2024).

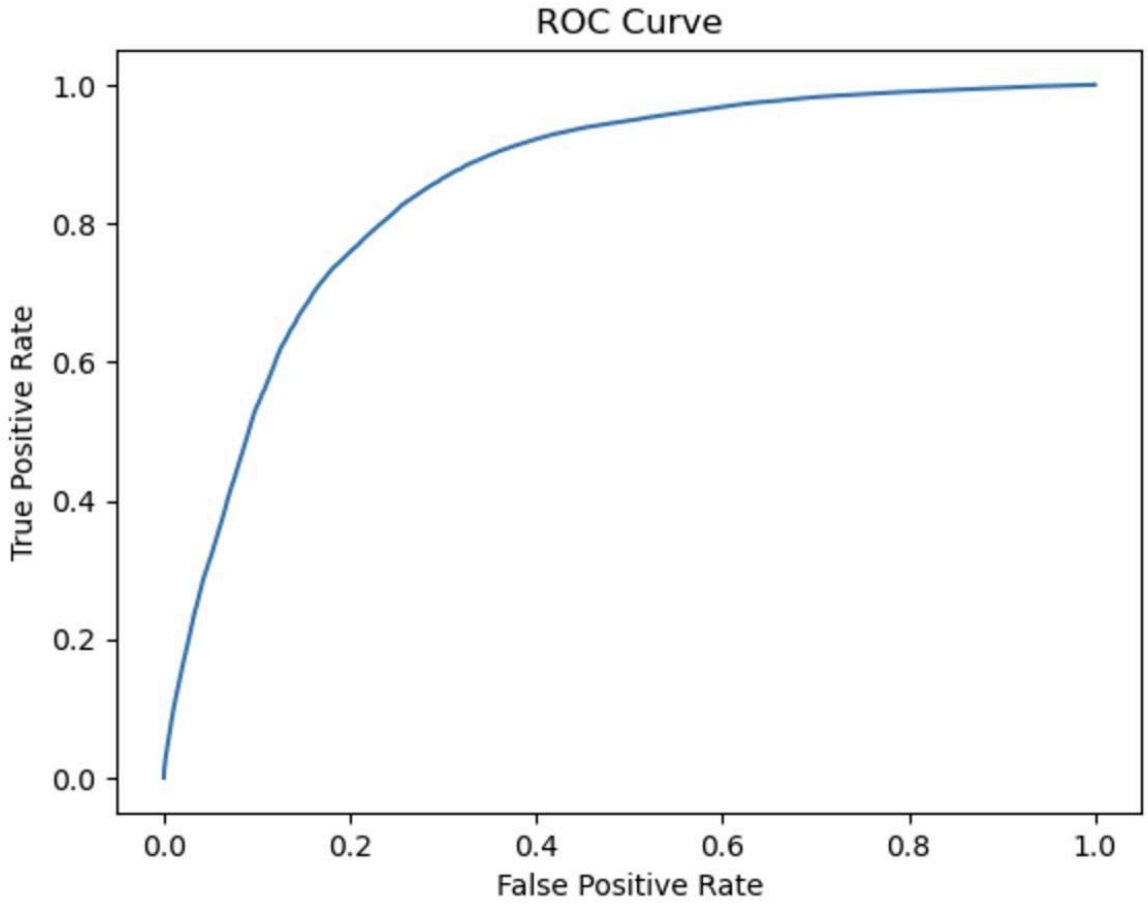
Appendix



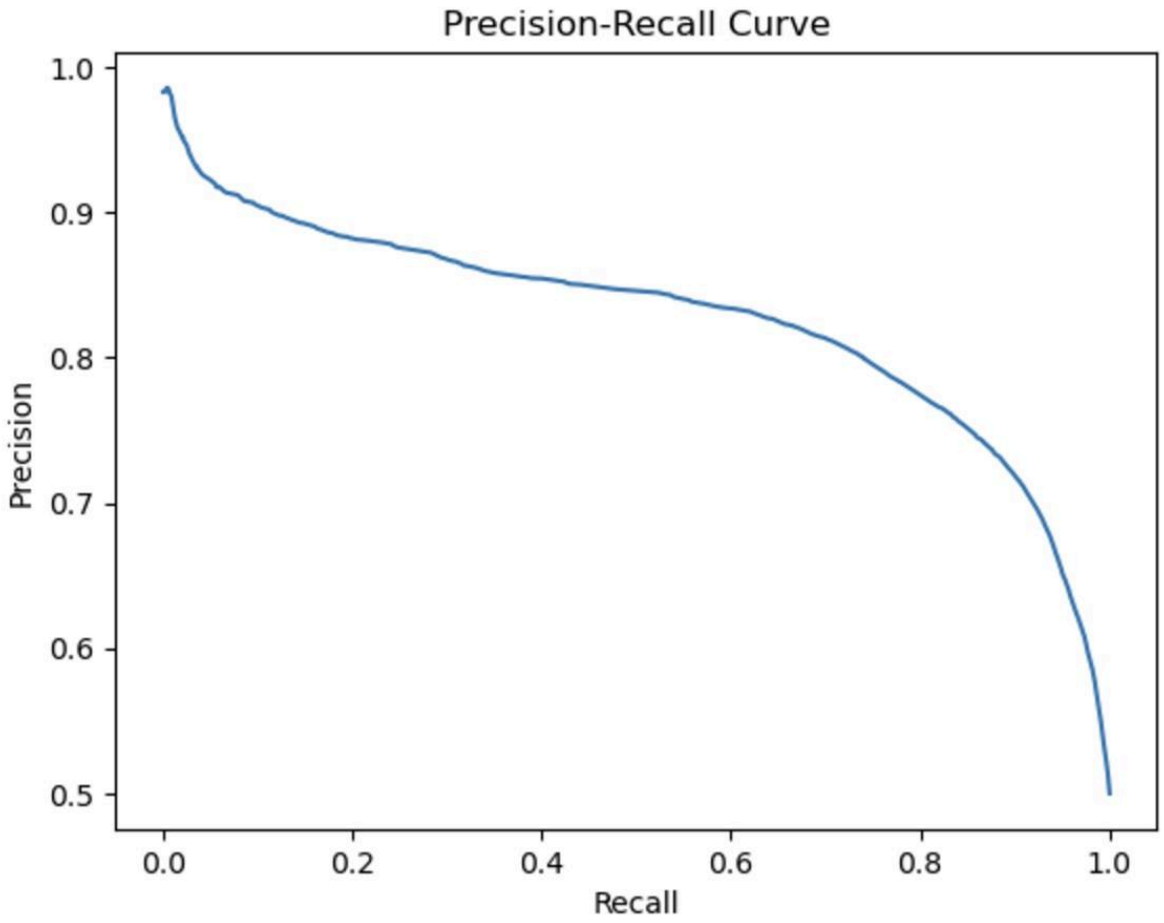
Scatter plot for actual victim_age vs predicted victim_age



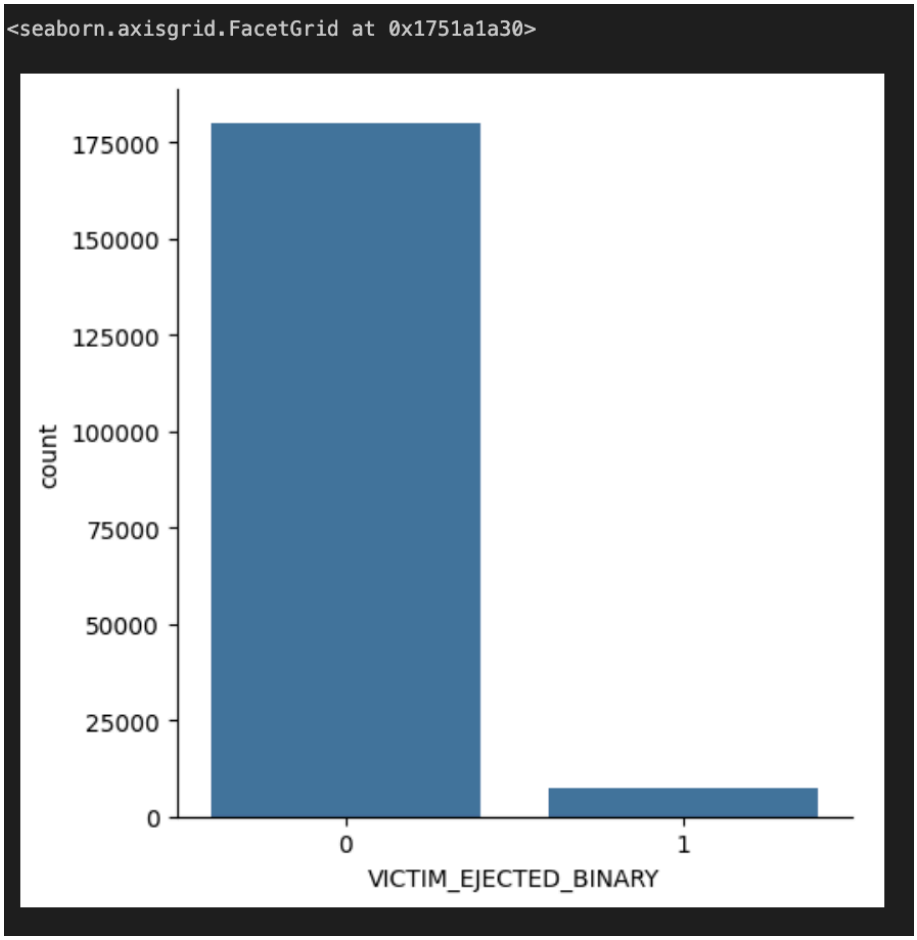
Logistic Regression (INJURY_SEVERITY_BINARY): Target Variable Imbalance



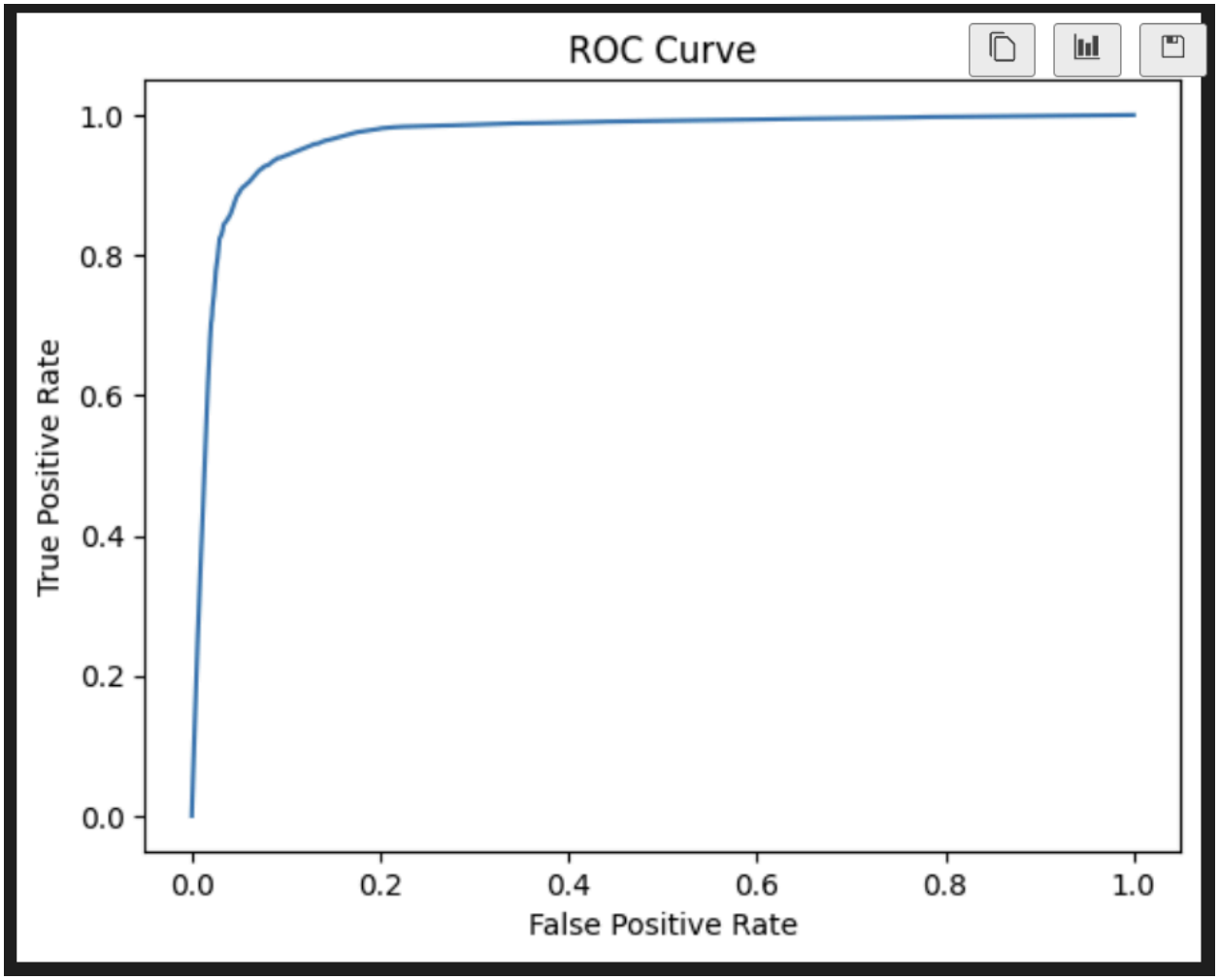
Logistic Regression (INJURY_SEVERITY_BINARY): ROC curve



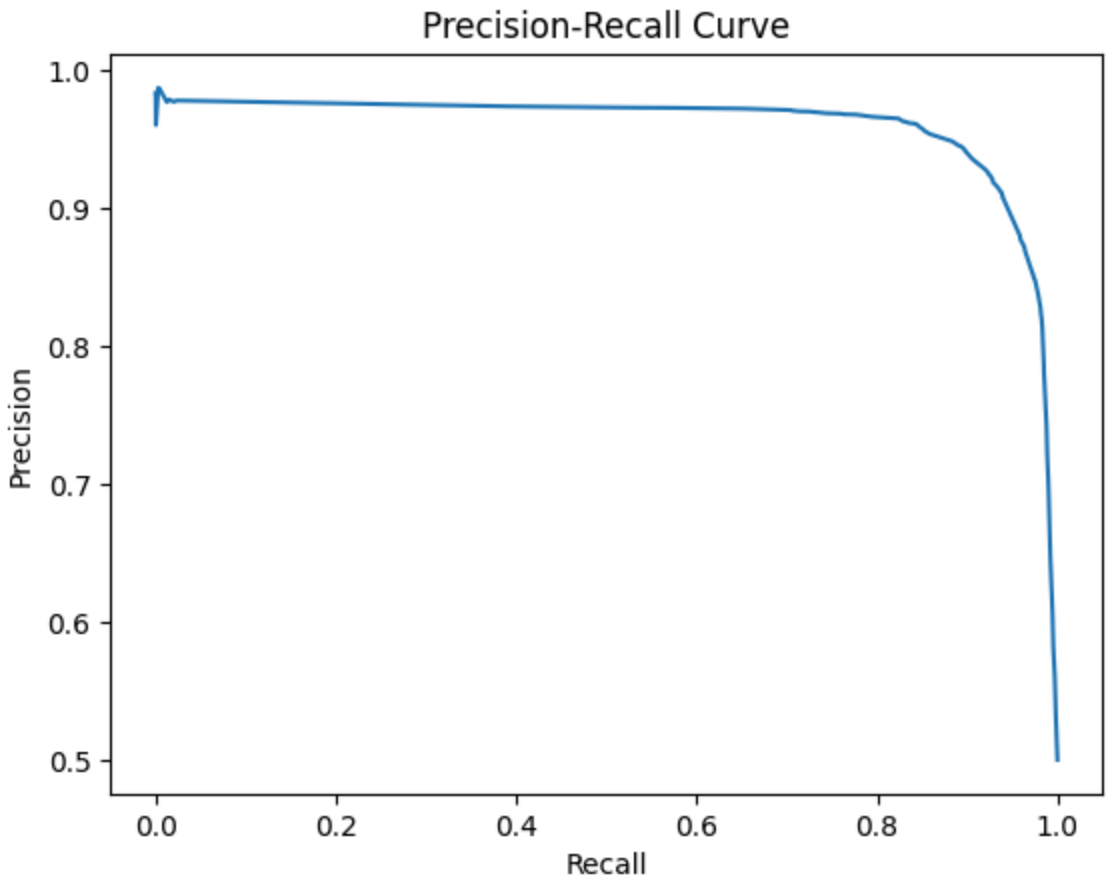
Logistic Regression (INJURY_SEVERITY_BINARY): Precision-Recall Curve



Logistic Regression (VICTIM_EJECTED_BINARY): Imbalance chart



Logistic Regression (VICTIM_EJECTED_BINARY): ROC Curve Chart



Logistic Regression (VICTIM_EJECTED_BINARY): Precision-Recall Curve Chart