TRAFFIC FLOW AND COLLISION PREDICTION IN LOSANGELES USING MACHINE LEARNING ALGORITHM

Manikanta Medidi,

Bapureddy Yenugula,

Teja Sai Kumar Nakka,

Sahan Vennam

Introduction

This research project aims to apply machine learning algorithms to improve traffic control and public safety in Los Angeles. The goal is to create a predictive model that uses the Los Angeles Traffic Collision dataset to rank places according to the frequency of collisions. This is related to previous research as the following analysis provides an effort that gives law enforcement priority over particular ranks. This helps in relation to the patrol deployment optimization. The suggested approach aims to create a thorough machine-learning model to help law enforcement make data-driven decisions and predict traffic density and crashes, enhancing public safety and urban mobility.

Goals and Objectives

Motivation

Overcoming the flaws in earlier research on traffic flow and accident prediction in Los Angeles is a critical issue that this effort attempts to address.

Significance

This Project of designing a model for predicting the number of collisions at different locations in Los Angeles, mainly aims for public safety, here this is the main thing that we are talking about and the law enforcement.

Objectives and Features:

It specifically aims to respond to the following questions:

1. How can the machine-learning model address the historical inconsistency in Los Angeles traffic location ranking?
2. In what ways can the model help law enforcement close the current gap in support for optimized patrol areas by helping them maximize patrol regions in real time based on traffic density?

3. How can forecast mistakes be reduced and traditional traffic pattern and congestion forecasts be superseded by utilizing machine learning approaches?
4. In what ways does the project's emphasis on traffic management and public safety address practical issues like improving urban traffic control?
5. What measures are implemented to guarantee that the machine-learning model that has been constructed takes into account all pertinent factors, offering law enforcement enlightening data and aiding in the optimization of patrol deployments for enhanced public safety and urban mobility in Los Angeles?
6. Which machine learning methods such as xgboost, robust regression, ridge regression and conditional random fields are best for predicting collisions and evaluating projected errors to evaluate traffic locations?

Related Work (Background)

Here, this project is mainly aims for the public safety that are mainly concerned about the traffic collision in Los Angeles city, the collisions that are occurring at the particular locations which leads to question the public safety has inspired us and motivated to choose this topic, and the model is designed to analyze the traffic collisions at certain areas and try to take required precautions and related to the law enforcement.

Dataset

For this analysis, we have taken the dataset that contains all the measures required for analyzing the number of collisions, and this is the dataset that we have taken through Kaggle, which is the website for the datasets, this dataset will act as the benchmark of out analysis. Here is the link for the dataset that we have used:

https://www.kaggle.com/datasets/cityofLA/los-angeles-traffic-collision-data

Figure 1: Dataset

Detail design of Methods

Approach

Our approach here is to concentrate on the public safety that are mainly crucial about for this we are going to start right from the data

collection to the model evaluation and further studies.

Model Selection

Here to implement this methodology we are going to use different models like XGBoost, Ridge Regression model, Robust Regression model, Elastic Net regression model, and going to compare those and used for analysis.

Implementation

1. Data Collection and Preprocessing

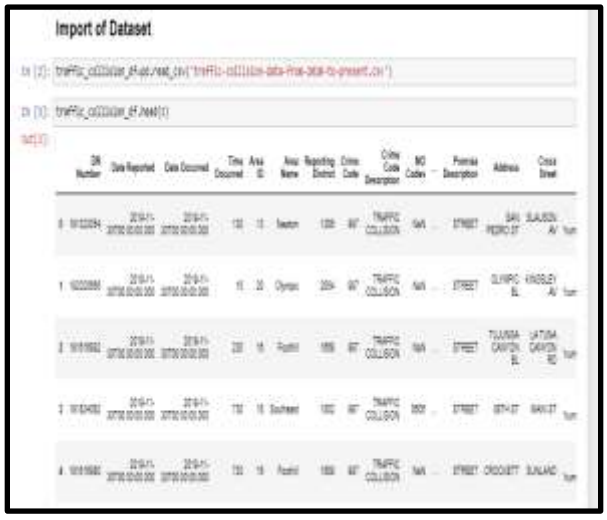

Figure 2: Importing the dataset.

(Source: Derived from Jupyter)

The following figure shows the import of the dataset into the Jupyter environment, which is a critical first step in the project. In order to conduct additional analysis, this procedure entails importing the Los Angeles Traffic Collision dataset into the computational workspace. The code snippets or instructions used to read and load the dataset are probably displayed in the figure. The basis for further stages like preprocessing, exploratory data analysis, and model creation is laid by a successful import of datasets. It is a crucial stage in the research process because it gives researchers access to and control over the data they need to train and assess the machine learning model.

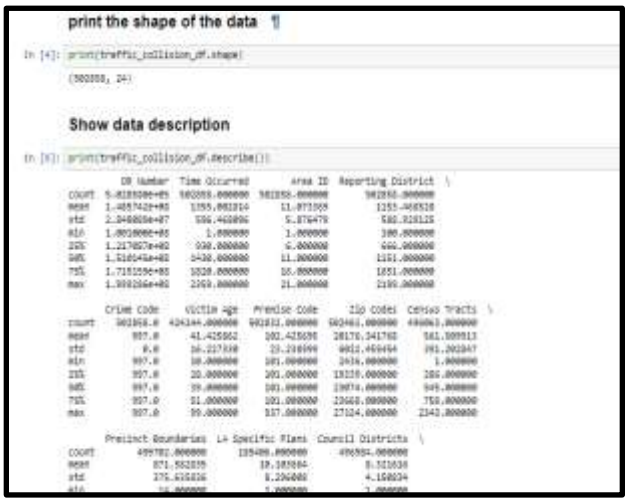

Figure 3: Printing the shape and showing the data description.

(Source: Derived from Jupyter)

Figure 2, "Printing the shape and showing the data description," offers an illustration of the most important discoveries found in the imported dataset. The dataset's shape, which is frequently shown as (rows, and columns), is printed to provide a brief summary of its dimensions and organization. This crucial data aids in comprehending the size of the

dataset and directs further analysis. The visualization of the data description is included in the graphic, providing statistical insights into different features. Important metrics for numerical qualities including mean, standard deviation, minimum, maximum, and quartiles are probably involved in this. These descriptive statistics give researchers a quick overview of the distribution and key tendencies of the dataset, making it easier for them to spot trends, anomalies, or possible research topics.

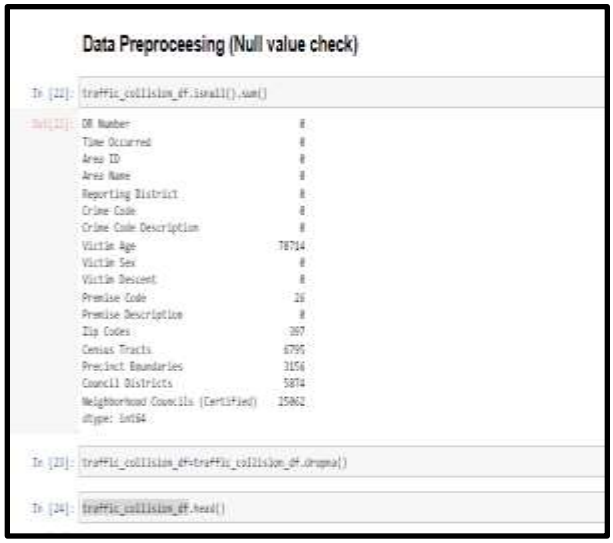

Figure 4: Checking and Removing Null Values

(Source: Derived from Jupyter)

The "Data Pre-process Null Check" picture probably shows a step in the data preprocessing process that deals with missing or null values in the dataset. Ensuring the quality and dependability of the data utilized for further studies and model training requires this procedure.

The code fragments or commands seen in the picture probably implement a null check on the dataset, locating and measuring any missing values in various columns. The findings might be shown graphically, showing whether or not each attribute has any null values. This is because null values can affect the precision and efficacy of machine learning models, handling them is an essential part of data preprocessing.

2. Exploratory Data Analysis (EDA)

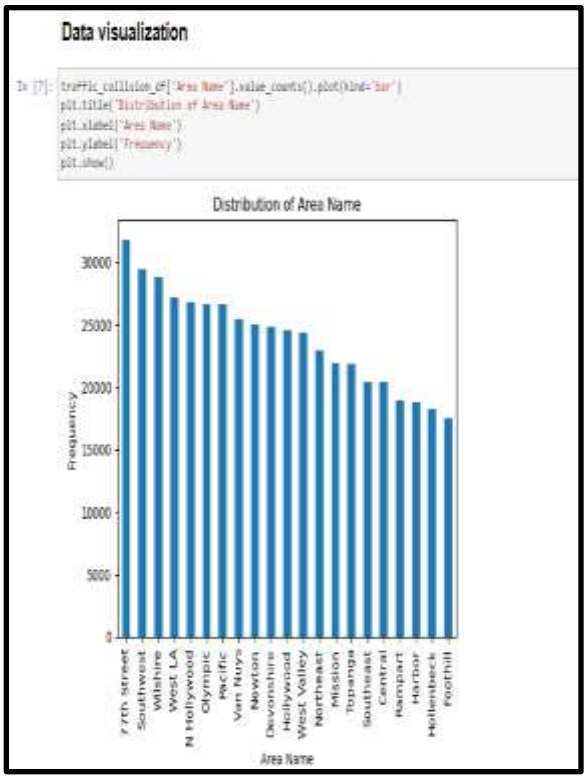

Figure 5: Visualization of Area Names

(Source: Derived from Jupyter)

Graphical representations demonstrating the correlation between traffic collision data and various Los Angeles areas are probably part of the "Visualization of the data from the perspective of area name in relation to Los Angeles Traffic Collision Data" process. Understanding trends and possible hotspots is made easier with the help of this visualization, which provides valuable insights into the distribution of collisions across different regions.

The Plots or charts show the frequency of crashes in various locations including a process of graphic to provide a visual comparison. Researchers could identify locations with greater or lower collision rates by, for instance, using a bar chart to show the total number of collisions in each area. The name of the x-axis is the area name and the name of the y-axis is frequency.

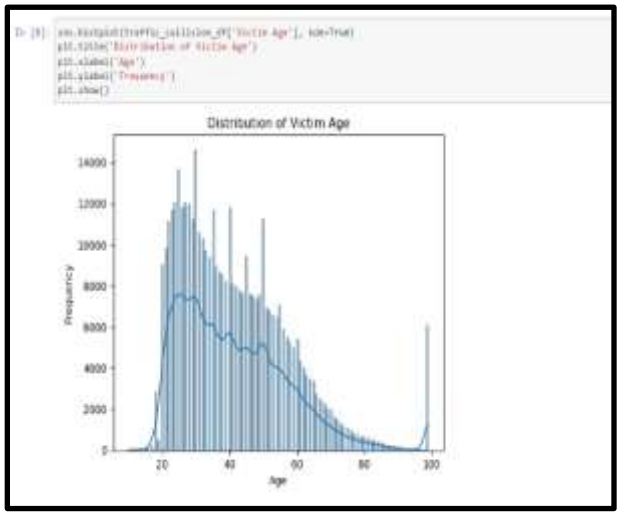

Figure 6: Visualization of Victim Age

(Source: Derived from Jupyter)

The picture that shows the age distribution of victims in traffic accidents in the US offers important information on the demographic trends in traffic accidents. The age group of 18 to 24 exhibits a significant peak, indicating that their members are more likely than members of other age groups to be involved in traffic collisions. This is consistent with the generally accepted notion that young adults—especially those between the ages of 18 and 24—tend to drive more recklessly. The slow decrease in collision frequency with age suggests that older age groups may have more experience and awareness when driving, or they may have changed their driving behaviors. In order to reduce unsafe driving, this graphic depiction emphasizes the significance of focused road safety projects and teaching programs directed at younger drivers.

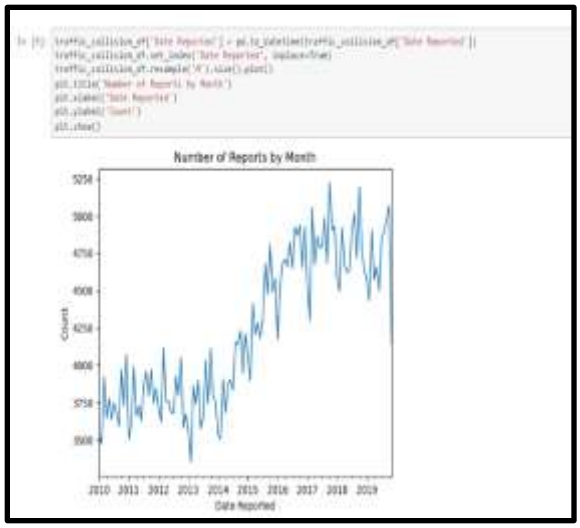

Figure 7: Visualization of the data of the number of reports by month

(Source: Derived from Jupyter)

The y-axis shows the number of reports of traffic collisions, while the x-axis shows the timeline in years, specifically from 2010 to 2019. The figure gives a clear and understandable summary of the monthly trends in traffic collision events by graphically representing the variation in the number of reports over time. Patterns, seasonality, or other significant changes in the frequency of recorded traffic crashes throughout the given time period can be found using this visualization technique. The resulting visualization provides insights into any long-term or seasonal trends by showing the amount of traffic collision reports over a period of months. The plot illustrates how the count varies across several months within the designated years. The x-axis shows the date reported, while the y-axis represents the count of reports.

3. Feature Selection

Figure 8: Performing Feature Selection and Splitting the Dataset

(Source: Derived from Jupyter)

The code snippet that is supplied focuses on feature selection for a machine-learning model that uses traffic collision data from Los Angeles. The variables "Victim Age," "Victim Sex," "Victim Descent," "Premise Code," "Premise Description," "Zip Codes," and "Census Tracts" are chosen to forecast the goal variable "Area Name." By concentrating on the variables that have the most influence, the model can improve its predictive skills. The characteristics that have been chosen in this particular context are intended to record crucial data about traffic collisions, including victim demographics, geographical specifics, and property details. The quality and applicability of the selected

features in precisely forecasting the target variable, "Area Name," based on the Los Angeles traffic collision data, are critical to the machine learning model's performance.

4. Model Selection

Figure 9: Selecting XGBoost Model

(Source: Derived from Jupyter)

The above code sample uses the potent machine learning method XGBoost to train a model using traffic collision data from Los Angeles. This is a broad explanation that is related to importing XGBoost as xgb: This command imports the XGBoost library from sklearn. metrics as 'xgb.' import error_squared: imports the sci-kit-learn library's mean squared error metric.

Figure 10: Selecting Ridge Regression Model

(Source: Derived from Jupyter)

In order to forecast the results of traffic collisions, Ridge Regression is used in the model selection process.

Figure 11: Selecting Elastic Net Regression Model

(Source: Derived from Jupyter)

The Los Angeles Traffic Collision Data is used to anticipate outcomes using the Elastic Net Regressor. Predictions are produced on the test set (X_test) after the model has been trained using the training data (X_train, y_train).

Figure 12: Selecting Robust Regression Model

(Source: Derived from Jupyter)

The robust regressor is used to model the results of traffic collisions. A particular epsilon parameter is used to instantiate this robust regression model. Predictions for the test set are produced once the model has been fitted to the training set. 5. Model Training and Hyperparameter Tuning

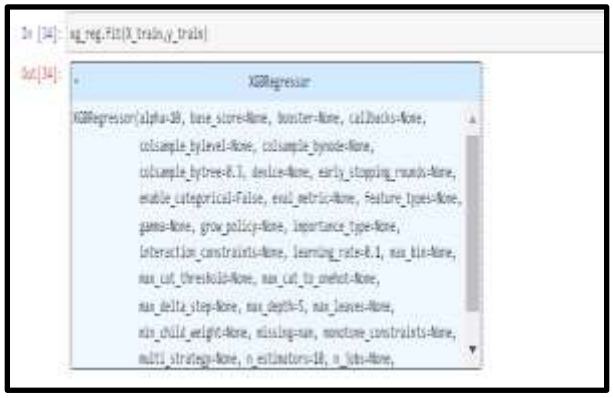

Figure 13: XGBoost Model Training

(Source: Derived from Jupyter)

The XGBoost Regressor model (xg_reg) that was trained on the Los Angeles traffic collision data is demonstrated in the provided code sample. The training set of data (X_train and y_train) is used to fit the model. The XGBoost Regressor has a number of hyperparameters set, including learning_rate, max_depth, and n_estimators.

Figure 14: Robust Regression Model Training

(Source: Derived from Jupyter)

The model's performance is evaluated by computing Mean Squared Error (MSE), and the result is reported. Los Angeles traffic collision data provides an example of how well the Huber Regressor, which is intended to manage outliers, predicts outcomes.

Figure 15: Elastic Net Regression Model Training

(Source: Derived from Jupyter)

The resulting figure, which shows real vs. projected values visually, sheds light on how accurate the model predicts values. Ridge Regression shows a Mean Squared Error of 35.94 and an accuracy of 6.07% in this case.

Results and Analysis

```
In [35]: y_pred = xg_reg.predict(X_test)

In [36]: mse = mean_squared_error(y_test, y_pred)
         print(f"Mean Squared Error: {mse}")

         Mean Squared Error: 18.317312490000586

In [37]: from sklearn.metrics import r2_score
         r2 = r2_score(y_test, y_pred)
         accuracy_percentage = r2 * 100

         print(f"Accuracy: {accuracy_percentage:.2f}%")

         Accuracy: 51.12%

In [38]: import matplotlib.pyplot as plt
         plt.figure(figsize=(5, 3))
         plt.plot(y_test, label='Actual')
         plt.plot(y_pred, label='Predicted', linestyle='dashed')
         plt.xlabel('Index')
         plt.ylabel('Value')
         plt.title('Actual vs Predicted')
         plt.legend()
         plt.show()
```

Figure 16: XGBoost Regression Model Evaluation

(Source: Derived from Jupyter)

The code uses traffic collision data from Los Angeles to assess an XGBoost Regressor's performance. To evaluate predictive accuracy, it computes the R2 Score and Mean Squared Error (MSE). The model's accuracy is indicated by the resulting MSE of 18.32. The explanatory strength of the model is indicated by its 52.12% R2 Score. By graphically comparing real and anticipated values, the plot sheds light on how well the model performed.

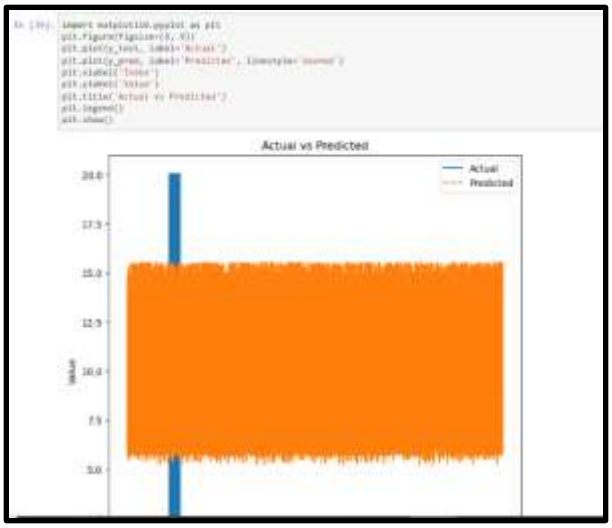

Figure 17: Actual vs Predicted Graph for XGBoost Model

(Source: Derived from Jupyter)

The Los Angeles Traffic Collision Data visualization shows real and anticipated data from an XGBoost Regressor model. The performance of the model is made easier to understand by the line plot, which illustrates the correlation between the actual and projected outcomes. A clear visual representation of the prediction accuracy of the model is given by dashed lines, which separate anticipated values from actual values.

Figure 18: Ridge Regression Model Evaluation

(Source: Derived from Jupyter)

This is for the purpose of making predictions using Los Angeles traffic collision data, the code uses Ridge Regression. The model yields a precision-indicating Mean Squared Error of 35.94. This is by visually comparing actual and anticipated values, the scatter plot provides information about the efficacy of the model.

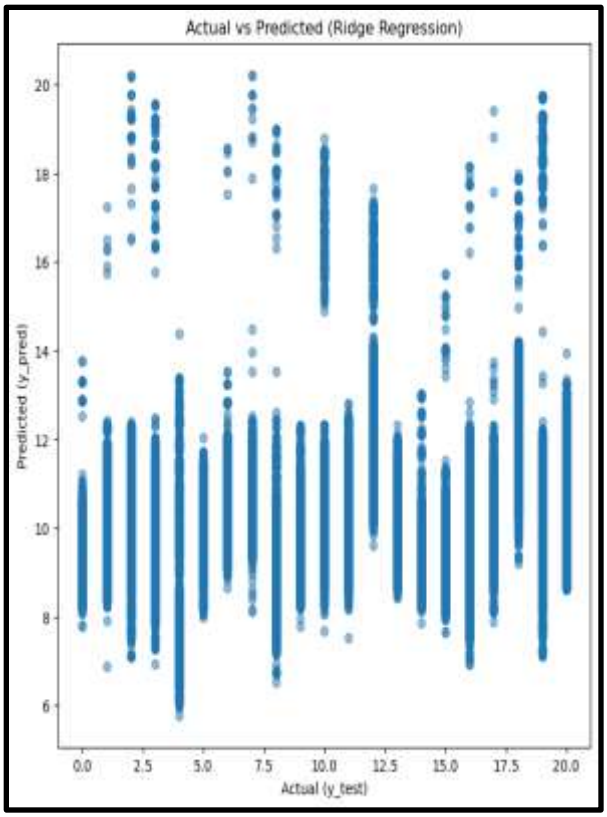

Figure19: Actual vs Predicted Graph for Ridge Regression Model

(Source: Derived from Jupyter)

The link between real and anticipated values in the Los Angeles traffic collision data from the Ridge Regression is displayed visually in the scatter plot. The model's efficacy is indicated by the points aligned around the diagonal, where actual results closely match forecasts. The graphic facilitates the evaluation of the Ridge Regression forecasts' dependability and correctness.

Figure 20: Elastic Net Regression Model Evaluation

(Source: Derived from Jupyter)

This is to predict results, the Elastic Net Regressor is applied to the Los Angeles Traffic Collision Data. Following the model's training with the training set (X_train, y_train), predictions are generated on the test set (X_test).

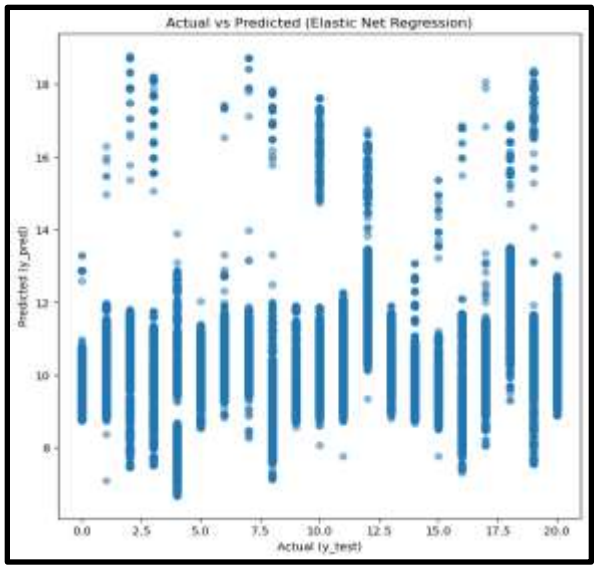

Figure 21: Actual vs Predicted Graph for Elastic Net Model

(Source: Derived from Jupyter)

The scatter plot visually assesses Elastic Net Regressor's accuracy by comparing actual and predicted values in Los Angeles Traffic Collision Data.

Figure 22: Robust Regression Model Evaluation

(Source: Derived from Jupyter)

In order to better handle outliers and improve model dependability in the Los Angeles traffic collision data, robust regression employing the Huber Regressor is used to estimate traffic outcomes.

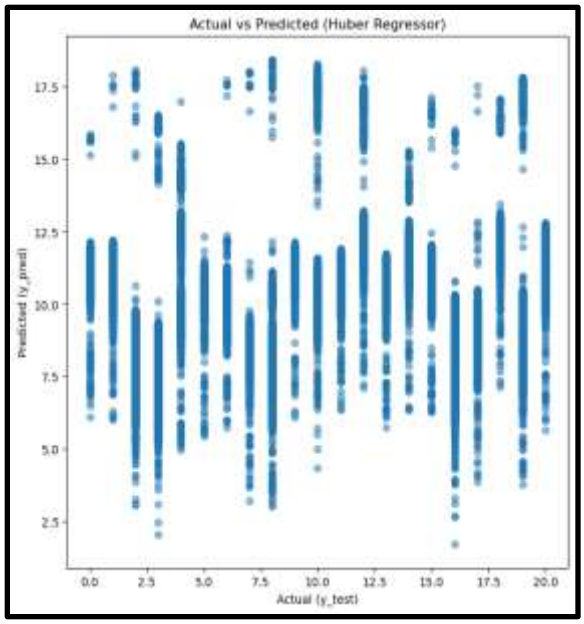

Figure 23: Actual vs Predicted Graph for Robust Model

(Source: Derived from Jupyter)

The graphic highlights the robust regression's performance in Los Angeles traffic collision data by showing the agreement between real and projected values and highlighting its resistance to outliers.

Project Management:

Coming to the management of the project, it has started with collection of the data, here we have chosen from the Kaggle website, and we have taken dataset of traffic collision data in Los-Angeles, Bapu Reddy Yenugula involved in data collection, describing, and preprocessing of that dataset, coming to the data visualization and Exploratory Data Analysis, it was handled by the Sai Kumar Nakka, next Model selection is been done by the Manikanta Medidi, here we have taken four models those are XGBoost, Ridge Regression, Elastic net Regressor, and Robust Regressor, by using these he has done the required analysis and Sahan Vennam has involved in making Model Evaluation, and making report is been done by four of us, as we all involved in it.

Deployment and Recommendations

Los Angeles should improve traffic control in locations where collision rates are expected to be greater, according to model forecasts. This is the purpose of planning the respective resource allocation that is making the most use of patrol deployments and putting specific safety precautions in place. Public safety should be greatly enhanced, and traffic accidents can be decreased with constant observation and modification of actions based on real-time forecasts.

Conclusion

The main goal of the project is to use various machine learning approaches to improve traffic management and public safety in Los Angeles. Forecasting and rating of collision-

prone sites are made possible by the created Machine Learning Model of Traffic Collision Prediction. This encourages efficient traffic management by enabling law enforcement to proactively reduce accident risks. The project's completion represents a useful instrument for proactive interventions, enhancing public safety by proactively tackling traffic issues in Los Angeles via data-driven insights and predictive analytics.

Implementation report

Work Completed

We have completed all the required analysis, right from data collection, cleaning, model selection, feature evaluation, model evaluation, deployment, and recommendation to the conclusion.

Description

Here we have done the project to build a model that is used to evaluate the collisions in Los Angeles city, and public safety is the main goal of this project and we have completed it.

Responsibility

Bapureddy Yenugula – Data Collection and Preprocessing.

Sai Kumar Nakka – Data Visualization and EDA

Manikanta Medidi – Model Selection and Analysis

Sahan Vennam – Model Evaluation

Contributions

Everybody has completed their allotted tasks in the clean and neat way, and this is the main reason for the completion of our project in a stipulated time interval and coming to the percentages, each has contributed one fourth of the project and it is equally divided as well.

Bibliography

Al-Qaness, M.A., Abd Elaziz, M., Hawbani, A., Abbasi, A.A., Zhao, L. and Kim, S., 2019, October. Real-time traffic congestion analysis based on collected tweets. In 2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS) (pp. 1-8). IEEE.

Ghafelebashi, A., Razaviyayn, M., Dessouky, M. and Center, M.T., 2021. Incentive Systems for New Mobility Services [Supporting Dataset] (No. NCST-USC-RR-22-26). METRANS Transportation Center (Calif.).

Ma, J., Ding, Y., Cheng, J.C., Tan, Y., Gan, V.J. and Zhang, J., 2019. Analyzing the leading causes of traffic fatalities using

XGBoost and grid-based analysis: a city management perspective. IEEE Access, 7, pp.148059-148072.

Pourpanah, F., Lim, C.P., Etemad, A. and Wu, Q.J., 2023. An Ensemble Semi-Supervised Adaptive Resonance Theory Model With Explanation Capability for Pattern Classification. IEEE Transactions on Emerging Topics in Computational Intelligence.

Ugirumurera, J., Gomes, G., Porter, E., Li, X.S. and Bayen, A.M., 2018, November. A unified software framework to enable solution of traffic assignment problems at extreme scale. In 2018 21st International Conference on Intelligent Transportation Systems (ITSC) (pp. 3917-3922). IEEE.

Yang, C., Chen, M. and Yuan, Q., 2021. The application of XGBoost and SHAP to examining the factors in freight truck-related crashes: An exploratory analysis. Accident Analysis & Prevention, 158, p.106153.