Road Accident Analysis And Traffic Severity Prediction Using Advanced Data Mining Techniques

Niharika Khanna, Arpita Rawat
Department of Artificial Intelligence and Data Science
Bhagwan Parshuram Institute of Technology
Rohini, Sector-17, New Delhi- 110089
{khanna.niharika09, arpitarawat984}@gmail.com

Dr. Varsha Sharma
Department of Artificial Intelligence and Data Science
Bhagwan Parshuram Institute of Technology
Rohini, sector-17, New Delhi- 110089
varshasharma@bpitindia.com

***Abstract*—Road accidents have a substantial economic impact, but their effects on lost lives are even more significant. In the USA alone, the National Highway Traffic Safety Administration released its latest projections for traffic fatalities in 2022, estimating that 42,795 people died in motor vehicle traffic crashes. Reducing these accidents is a challenging task, as evidenced by the numerous articles discussing the alarming rates of deaths in road accidents. This research project aims to predict traffic accident severity using advanced data mining techniques, focusing on factors like weather conditions, distance, and time of day. By analyzing and predicting accident severity, this study seeks to contribute to the development of effective strategies for reducing the detrimental effects of road accidents and saving lives.**

***Index Terms*—Crashes, Hot spots, Machine learning, GIS, Traffic accidents, Accident severity, Artificial intelligence, Predictive modeling, Traffic accident duration, Freeways, Accident factors, Severity prediction**

I. Introduction

The issue of road safety is increasingly gaining prominence as a significant societal issue globally. Recognizing the primary causes of road traffic accidents is critical for developing effective solutions to lessen the detrimental impact on human lives and property. Road severity is not random; it follows predictable patterns that can be predicted and minimised.

Accurate traffic severity predictions can assist in reducing response times of emergency services and improving overall road safety. This research paper aims to predict the severity of traffic accidents based on various features such as weather conditions, distance, and time of day. By leveraging advanced data mining techniques, the study seeks to uncover the key factors contributing to road accident severity and develop predictive models to support policy making and infrastructure improvements.

The motivation behind this research is the alarming statistics on road accidents and their devastating consequences. In the USA alone, the National Highway Traffic Safety Administration released its latest projections for traffic fatalities in 2022, estimating that 42,795 people died in motor vehicle traffic crashes. Reducing these accidents and mitigating their severity is a pressing challenge that requires a data-driven approach.

This research paper will explore the application of various machine learning algorithms, including decision trees, random forests, logistic regression, and neural networks, to predict the severity of road accidents. The study will also investigate the importance of different features in determining accident severity, providing valuable insights for targeted interventions and policy decisions.

By addressing the critical issue of road accident analysis and severity prediction, this research aims to contribute to the ongoing efforts to enhance road safety and save lives. The findings of this study can inform the development of early warning systems, optimize emergency response, and guide infrastructure planning and design to create safer roads for all users.

II. Literature Survey

In this literature survey, we review various research papers related to Traffic Severity Analysis that utilize different models for predicting accident severity.

2.1. Improved Naive Bayes Classification Algorithm for Traffic Risk Management The paper introduces the Naive Bayes classification method and its advantages in estimating necessary parameters based on limited training data. To address shortcomings of Naive Bayes, the authors propose an ”Improved Naive Bayes Classifier” by incorporating feature weighting and Laplace calibration to enhance accuracy.

2.2. Traffic Accidents Severity Prediction using Support Vector Machine Models This study explores the use of Support Vector Machine (SVM) models to predict accident

fatality rates, comparing radial basis function and linear kernel functions. The research focuses on accidents in Lebanon, preprocessing data by normalization and outlier removal. SVM aims to maximize the margin of the hyperplane for effective classification.

2.3. Modeling Road Accident Severity with Logistic Regression The paper discusses the application of Logistic Regression (LR) in analyzing traffic accident severity. LR helps identify factors correlated with accident severity, utilizing IBM Modeler 18.0 software for model training. The study categorizes accidents as ”serious” or ”minor” and emphasizes the importance of variable significance and feature selection.

2.4. Traffic Accident Analysis Using Decision Trees and Neural Networks This research employs Decision Trees and Neural Networks to analyze traffic accidents, focusing on variables correlated with severity. The study uses data from the Nigeria Road Safety Corps, training decision trees on categorical data and neural networks on continuous data. Results show the Decision Tree model outperforming Neural Networks, with specific emphasis on the Radical Basis Function (RBF) Neural Network's performance metrics.

These studies highlight the diverse approaches and methodologies used in predicting traffic accident severity, showcasing the importance of advanced data mining techniques in enhancing road safety and accident prevention strategies.

III. Dataset

A. Dataset Details

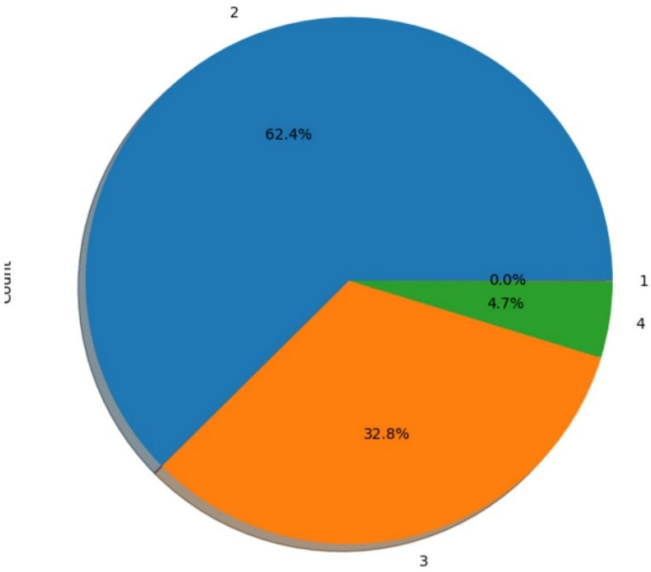
We used a countrywide traffic accident dataset available on Kaggle. It comprises of seven years of data, about 7700000 rows, and 46 columns. Since this is raw data, we would need to process and clean this data, and hence Pre-processing is very much needed. In the dataset, the traffic impacted due to accident, data were collected from February 2016 to March 2023, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset currently contains approximately 7.7 million accident records.

The filtered dataset consists of only the following columns: Year, Severity, Start Lat, Start Lng, Distance(mi), Street, City, County, State, Airport Code, Temperature(F), Wind Chill(F), Visibility(mi), Wind Direction, Weather Condition, Traffic Signal, Sunrise Sunset, TimeDiff

B. Data Pre-processing Techniques

We used the following pre-processing techniques to process raw data:

1. Handling missing values We checked for all NULL value entries in our dataset, which were around 10,000 in total, and deleted all such entries.



Percentage severity distribution

2. Handling duplicate values Our dataset contained around 5,000 repeated entries. We deleted all duplicates to make all rows unique.

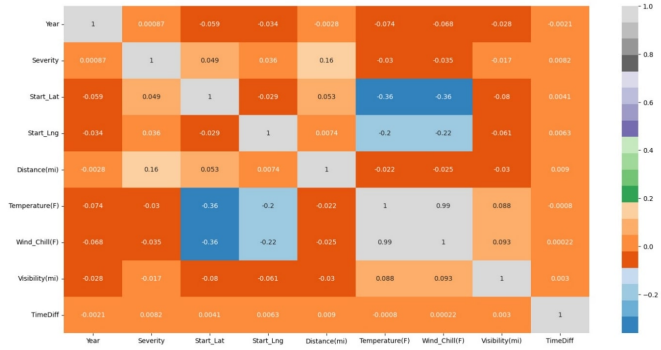
3. Slicing the dataset Our dataset initially contained about 7 million entries from 2016 to 2023. Training any model on such a large database is not time and resource-feasible. So, we only took entries from 2016 to 2018, bringing down the number of rows to around 3,00,000.

4. Encoding categorical variables Our cleaned dataset contained 9 numerical columns (type : float64, int64) and 9 categorical columns (type : object, boolean), which had to be encoded before applying any models on it. We used both Label Encoding and One-Hot Encoding in order to do so.

5. Splitting dataset into Training and Testing set We divided the number of entries into Training and Testing sets in the ratio 80 : 20.

6. Feature scaling We scaled the features in our dataset to the same range so no feature dominates over the other. We used Standardization using the StandardScaler class of sklearn.preprocessing library in order to do so.

C. Data Inferences



Corelation Heatmap

The pie-chart of the percentage severity distribution tells us that most of the traffic observed on the roads is of severity

level 2 (62.4%) and severity level 3 (32.8%). Traffic severity levels of 1 and 4 are rarely observed.

The correlation heatmap of our dataset shows the relationship between different pairs of features. For example, we observe that Wind Chill(F) and Temperature(F) are strongly positively correlated, TimeDiff and Severity are mildly positively correlated, and Temperature and Start Lat are moderately negatively correlated.

The bar graph depicting sunrise and sunset times indicates that the majority of accidents occur during daylight hours.

The bar graph representing weather conditions reveals that the majority of accidents occur during clear or overcast weather conditions.

IV. Methodology, Model Details

We have experimented with the following machine learning models:

A. Mixed Naive Bayes

Naive Bias is a supervised learning classification model. It uses the naive bayes formula with a naive bias assumption that data features are independent of each other 1. Algorithm We have made a custom class to handle Na¨ıve Bias Classification (NBC) and used scikit learn for testing purposes such as f1 score from sklearn.metrics.

• Upon running train on the initiated class for NBC, the model counts the required parameters conditioned on each value of the output and stores them for categorical columns and stores the relevant values of mean and std for Numerical data.

• Predict function takes in the row for which prediction is to be performed

• It checks which category does a particular column lie in. If it's a category then na¨ıve bias is applied using na¨ıve formula along with a Laplacian method with alpha =5 . If its numeric category then it applies the relevant Gaussian model.

• Predict alpha takes a custom alpha to find the prediction

• Accuracy score takes in test set and its ground truth and for each row in set runs predict function on it to give the accuracy score. Accuracy score alpha helps in finding the best fit alpha

• Predict weighted makes an attempt towards weighted na¨ıve bias

B. Support Vector Machine(SVM)

SVMs are supervised learning models with associated learning algorithms that analyze data for classification, regression and clustering analysis. We will be using it for the classification of traffic severity levels (In the range between 1-4).

1. Algorithm

• Need to determine kernel that performs best on the dataset.

• Determine best by hyperparameter tuning for a SVM classifier using grid search.

• Higher degree models can be likely to be overfit, used regularization,and likely prevent the model from overfitting the data.

• After performing a grid search, we will get the best-performing SVM model along with regularization parameters and therefore we can train the best-performing model on training data.

• We will then check its performance on test data to know the performance of model.

• We use the cross-validation technique to determine the performance of model

C. Logistic Regression

1. Algorithm

• The first model runs on the dataset with only 2-3 severity rating as these two are the major severity values.

• The second model is a multilevel logistic regression model, where the data is first classified into more and less severe categories.

• Less severe data is then classified into 1 and 2 categories of severity, while more severe data is classified into 3 and 4 categories of severity.

D. Decision Trees and Boosting

1. Algorithm

• The model reads the input data upon calling the fit function

• If it is at the node, then split is made such that Info Gain is the highest. Else, it is at the leaf it checks if all the data is being classified properly or not

• If the classification is incorrect, then the algorithm makes a split is continued if termination condition is not being fulfilled

• Algorithm stops on reaching the termination condition

2. Algorithm Random Forests

• The model reads the input data upon calling the fit function and sees the number of decision trees to be made(n)

• The model proceeds to make a decision tree and randomly chooses m features to be used for building the tree and train data is created by means of bootstrapping.

• In the end, the model ends by making n different decision trees, each trained on a different bootstrap data and made by splitting on m randomly chosen features

3. Boosting Algorithm

• Ada Boosting - This works by creating decision stumps(depth=1) and having equal weights for all, the misclassified examples are given higher weightage as the algorithm progresses

• Gradient Boosting - The primary aim of this boosting method is to decrease the loss function

• Xtreme Gradient Boosting - It is a modified version of Gradient Boosting which gives higher weights and minimizes the loss along with the use of parallelization and cache

E. Multi-Layer Perceptron (MLP)

MLP is a type of artificial neural network that consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function.Its used for both classification and regression tasks

1. Algorithm

• The MLP model was implemented using the MLPClassifier from the sklearn.neural network library.

• Grid search was performed using different activation functions, including 'tanh', 'relu', 'logistic', and 'identity' on subset dataset of 5k entries.

• The best parameters from the grid search were then used to initialize the model.

• The model with the highest accuracy on the training set was selected as the best model and then trained individually on the complete dataset.

• Using the best-performing activation function, the model was learning and improving its performance on the training set and had better accuracy scores on both the training set and test set.

V. Results And Analysis

We implemented the following models: Naive Bayes Classifier, which makes use of Laplace and weighted cor5 rection, Support Vector Machine, which takes in various kernel functions like linear and Radial Basis Function(RBF), logistic regression classifiers where the first classifier is a simple logistic regression and the other is a multilevel logistic regression, Decision Trees and Random forests, boosting algorithms like Gradient Boost and XGBoost, and finally a Multi-Layer Perceptron.

The reason why linear regression is not used is that severity can take up only 4 values which are discrete values, while linear regression works best for predicting real numbers given the parameters hence, we chose models like Naives Bayes, SVM and Decision Trees, which give discrete values. SVM's main use is to classify binary data, but it works well on multiclass data. This is possible as scikit-learn's implementation of SVM considers 2 classifying factors, whether it is a part of a severity class or not. In other words, it breaks down the data internally into binary classes.

We also performed K-means clustering and the KNearest Neighbors Algorithm on our dataset. However, both these gave low accuracy scores because our dataset can't be clustered properly.

It is observed that the XGBoost gave the best accuracy of 0.913 on the testing test, followed by Random Forest with a high accuracy of 0.885, Gradient Boost with an accuracy of 0.858, and Decision Tree also having a high accuracy of 0.83. Apart from these, Support Vector Machine and Mixed Naive Bayes also gave a decent accuracy of 0.681 and 0.66, respectively.

VI. Conclusion

1. In this report, we explored the prediction of traffic severity using machine learning models and tried to analyze the dataset with various techniques to determine the best models.

2. Among classification models, Random Forest demonstrated very high accuracy among the models tested, achieving 89% accuracy on the test dataset.

3. Algorithm boosting algorithms trained on different decision trees also gave great results. For example, we were able to achieve 91% accuracy by using XGBoost.

4. We will need to save and restore/reload later our ML Model so as to test our model with new data or to compare multiple models or anything else. Hence, serialization and deserialization of models are required, which we will complete before the final evaluation of all the final models.

5. A real-world applicability is that with such high accuracy, these models could be potentially used in realworld applications, such as in traffic management systems, to predict and manage traffic severity which would lead to a better quality of life.

References

[1] Hong Chen, Songhua Hu, Rui Hua, and Xiuju Zhao. Improved naive bayes classification algorithm for traffic risk management. Journal on Advances in Signal Processing, 2021.

[2] Mu-Ming Chen and Mu-Chen Chen. Modeling road accident severity with comparisons of logistic regression, decision tree and random forest. 2020.

[3] Zeinab Farhat, Ali Karouni, Bassam Daya, Pierre Chauvet, and Nizar Hmadeh. Traffic accidents severity prediction using support vector machine models. International Journal of Innovative Technology and Exploring Engineering, 2020.

[4] Victor Olutayo and Adekunle Eludire. Traffic accident analysis using decision trees and neural networks. International Journal of Information Technology and Computer Science, 6:22–28, 01 2014.