ML Project Report Group 77

Anonymous submission

Paper ID

Abstract

Road accidents are a significant public health concern, and understanding the factors contributing to their severity is crucial for effective prevention strategies. Machine learning provides a powerful tool to analyze vast and complex accident datasets, uncovering hidden patterns and correlations. By accurately predicting accident severity, these models can help authorities prioritize interventions, optimize resource allocation, and implement proactive safety measures, ultimately leading to safer roadways and reduced accident impact.

The source code is available at: GitHub Repository

1. Introduction

The aim of our project is to determine the main factors influencing the severity of accidents. We'll develop a model that predicts accident severity, ranging from 1 to 4, with 1 representing minimal traffic disruption and 4 indicating a major impact.

In particular, the model will predict whether an accident is likely to be severe or not, without needing detailed information such as driver or vehicle data. It will be capable of processing real-time traffic accidents and providing real-time predictions for severe accidents.

2. Literature Survey

2.1. Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity

This study evaluates the performance of four machine learning techniques—Random Forest (RF), Logistic Regression (LR), Naïve Bayes (NB), and AdaBoost—in predicting traffic accident severity. The performance metrics used for comparison include precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC), all of which were derived from confusion matrices. The results indicate that Random Forest consistently outperforms the other models, achieving the highest accuracy (75.5%) and AUC value, highlighting its robust classification capability. RF's ability to handle both discrete and continuous data, along with its resistance to noise, contributes to its superior performance in accident prediction.

This research supports the application of RF in traffic safety systems, providing valuable insights for road design and accident prevention. However, the study acknowledges the limitation of missing data on key factors such as driver and pedestrian characteristics, which may impact accident severity. Future research should address these data gaps for more comprehensive modeling.

2.2. Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique

This study examines four distinct models: Feedforward Neural Networks (FNN), Support Vector Machines (SVM), and their respective fuzzy c-means clustering variants (FNN-FCM, SVM-FCM). Utilizing a dataset comprising 10,000 traffic crash incidents, the research emphasizes the importance of systematic parameter optimization to enhance classification accuracy. For instance, the FNN model achieved the best performance with the Levenberg–Marquardt training algorithm, employing an architecture of two hidden layers (32 and 2 neurons) and utilizing hyperbolic tangent sigmoid and softmax activation functions. This architecture led to an optimal learning rate of 0.00002 and a target loss of 0.000001.

In contrast, the SVM model incorporated a Gaussian radial basis function kernel, with careful tuning of parameters C=15 and =150 to balance training error against model complexity, achieving an accuracy of 73% Furthermore, the integration of fuzzy c-means clustering significantly influenced model performance. The optimal number of clusters was determined through iterative testing, yielding improved accuracies of 71.8% for FNN-FCM and 74.2% for SVM-FCM. This indicates that leveraging clustering not only refines input data classification but also facilitates better generalization in the prediction model, demonstrating the effectiveness of hybrid algorithms in machine learning applications for traffic safety analytics.

2.3. Accident Severity Prediction Using Machine Learning

Kumar and Santosh compared three models—Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR)—using attributes such as traffic and weather. Random Forest outperformed the others with 74% accuracy, while Decision Tree and Logistic Regression achieved 67% and 62%, respectively.

The superior performance of RF is due to its ensemble approach, making it more resistant to noise and better at handling diverse data types. Comparatively, studies by Chen et al and Satu et al. reported similar results with RF and DT, showing accuracy around 73%. The slight improvement in Kumar and Santosh's work is attributed to using additional contextual features.

3. Dataset and Pre-Processing

We used the US Accidents dataset by Sobhan Moosavi (published here: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents). The accident data was collected from February 2016 to March 2023, across 49 states in the US, using multiple APIs that provide streaming traffic incident (or event) data. The dataset currently contains approximately 7.7 million accident records and has 46 columns.

3.1. Description of the Dataset

The data consists of 46 attributes:
1) The severity of the accident.
2) Location of the Accident with start and end coordinates, street, city, county, state, country, zip code, airport code, etc.
3) Weather conditions like wind speed, temperature, humidity, pressure, visibility, weather condition(text), etc.
4) Time of the accident i.e. start time, end time including information about sunset and sunrise.
5) Road descriptions during the accident like a bump, crossing, junction, amenity, railway, roundabout, stop, traffic signals etc.

3.2. Exploratory Data Analysis (EDA)

List of EDA Performed:
1) Distribution of Severity.
2) Correlation Analysis for Numerical Features.
3) Distribution of Accidents by City.
4) Impact of Junctions on Severity.
5) Scatter Plot: Distance vs Severity.

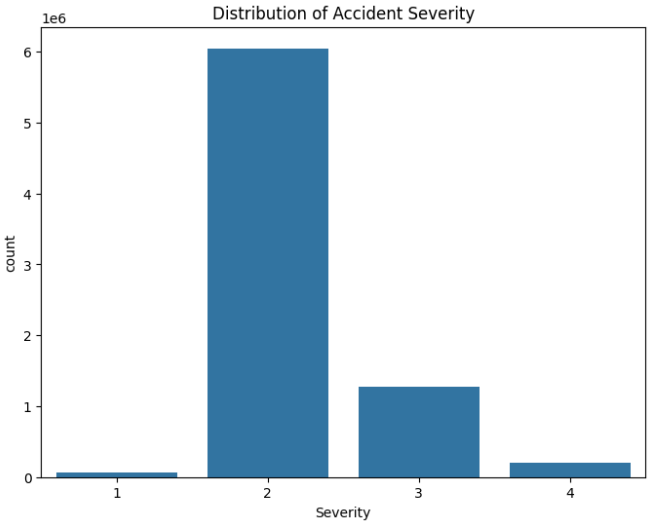


Figure 1. Distribution of Severity

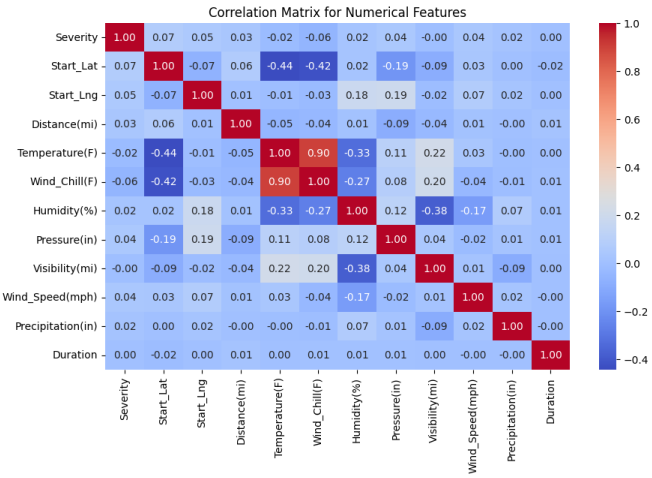


Figure 2. Correlation Analysis for Numerical Features

3.3. Processing the Dataset

1) We checked the number of false values in the boolean columns and dropped the following columns with the majority of entries as false:
Amenity, Bump, Give_Way, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Turning_Loop.
2) We decided to drop the following columns as they were closely related to the column Sunrise_Sunset:
Nautical_Twilight, Astronomical_Twilight, Civil_Twilight.
3) We checked the number of null values in all of the columns and found that the majority of entries in End_Lat, End_Lng were null, so we dropped both columns.
4) We dropped the Country column since it was a uniform column displaying US.
5) We removed columns like Airport_Code, ID,

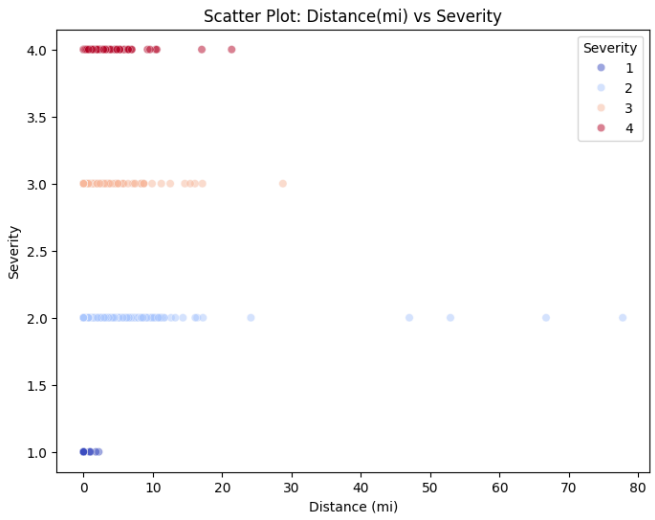


Figure 3. Scatter Plot: Distance vs Severity

Weather_Timestamp as these were irrelevant to our analysis.

6) We mapped the true and false values in boolean columns, i.e., Junction, Crossing, Traffic_Signal to 1 and 0, respectively.

7) We replace the null values with mean in the following columns: Wind_Chill(F), Wind_Speed(mph), Visibility(mi), Precipitation(in), Pressure(in), Temperature(F), Humidity(%) .

8) We did binning on the Start_Time into Year, Month, Day, and Time_S. We further binned Time_S to TimeofDay i.e. morning, afternoon, evening, and night. We further binned Month into Seasons i.e. 'Spring',' Summer', 'Autumn' and 'Winter'. We binned Day into Day_Type, i.e. Weekday and Weekend. We then dropped the original columns.

9) We dropped the columns where the entries in Weather_Condition, Weather_Direction are null, as these are important features.

10) There were two columns with categorical values, Wind_Direction and Weather_Condition. We simplified Weather_Direction into 11 unique classes by merging similar values. Weather_Condition contained 100+ unique values, out of which many values were closely related to each other. We grouped these values into 10 categories 'Clear', 'Cloud', 'Rain', 'Heavy Rain', 'Snow', 'Heavy Snow', 'Fog', 'Dusty', 'Windy', 'Ash'.

11) We finally settled on using 28 columns as our features: Severity, Distance(mi), Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Direction, Wind_Speed(mph), Precipitation(in), Clear, Cloud, Rain, Heavy_Rain, Snow, Heavy_Snow, Fog, Dusty, Windy, Ash, Junction, Crossing, Traffic_Signal, Sunrise_Sunset, TimeofDay, Season, Day_Type, Duration.

4. Methodology and Model Details

We conducted a thorough examination of the dataset to identify missing values and inconsistencies. Records with substantial missing data or anomalies were either imputed or removed. Following comprehensive exploratory data analysis (EDA) and a review of relevant literature, we identified 18 features that were deemed irrelevant to our analysis; these were subsequently removed to reduce the dimensionality of the model.We standardized and normalized the features to ensure that all remaining variables contributed equally to model training and applied log transformation to Duration and Precipitation to reduce skewness.

As an initial step, we conducted a baseline model experiment using LazyPredict on the dataset to establish a reference point for performance.

We then trained a Random Forest model to assess each feature's importance and identified the ten most important features. Additionally, we calculate mutual information scores between the features and the target variable to uncover any non-linear dependencies, further helping to identify the most relevant features. Combining the top features from both methods, we select a refined set of features for the final model.

The data set exhibited significant class imbalance. The distribution of the sample points was significantly skewed toward classes 2 and 3. This severe skew in the distribution posed a challenge for the classifier, as it could potentially bias the model toward the majority classes while neglecting the minority class. To combat this issue, we implemented SMOTE (Synthetic Minority Oversampling Technique). This technique generates synthetic samples for the minority class by interpolating between existing samples, creating a more balanced data set.

We performed hyperparameter tuning on the top-performing models to optimize their performance. This involved systematically exploring combinations of key parameters, such as the number of estimators, maximum depth, and learning rate, using techniques like grid search or random search. By fine-tuning these parameters, we aimed to achieve the best possible results while ensuring the models were not overfitting or underfitting the data.

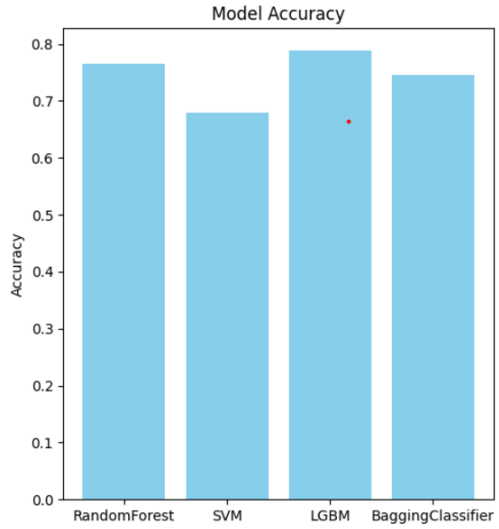
5. Results and Analysis

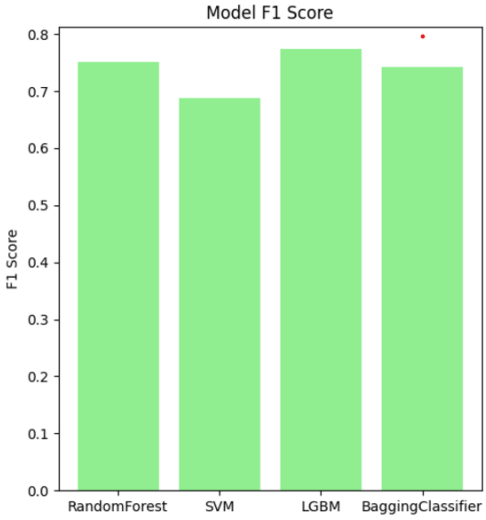
Through our analysis, we identified the following features as the most important in predicting the accident severity: **Wind_Chill(F)**, **Distance(mi)**, **Humidity(%)**, **Duration**, **Wind_Speed(mph)**, **Start_Lng**, **Start_Lat**, **Temperature(F)**, **Pressure(in)**, **Precipitation(in)**.

Surprising **Time of Day** and **Weather conditions** had little impact on predicting accident severity.

By applying SMOTE (Synthetic Minority Over Sampling Technique), selecting the key features, and fine-tuning

the hyperparameters to identify the optimal parameters, we obtained the following performance on select models:





LGBM Classifier performed the best among the models with Random Forest and Bagging Classifier not far behind in performance.

6. Conclusion

The majority of the models we trained showed a bias towards a single class. To address this, we applied several techniques, including SMOTE and preprocessing, which should contribute to improved results. Features like Humidity, Visibility, and Precipitation were found to be the most influential in the model's performance.

These features play a critical role in enhancing the accuracy and reliability of the predictions, as they provide significant insights into the underlying patterns in the data.

The findings from this study can guide the development of safety measures and strategies aimed at minimizing accident severity. In future work, the model can be further enhanced by incorporating additional features, including more detailed traffic data, and experimenting with alternative algorithms to boost predictive performance.

Throughout this project, we gained valuable experience in various areas of data science and machine learning. We learned how to handle large datasets efficiently using **Polars**, a fast and memory-efficient alternative to Pandas, which allowed us to perform data manipulation with improved performance. Our teamwork and collaboration skills were strengthened through the use of tools like **Git** for version control and effective communication within the team. Additionally, we applied **SMOTE** (Synthetic Minority Over-sampling Technique) to address class imbalance in the dataset, enhancing the model's ability to predict minority class outcomes. We also focused on data preprocessing, where we analyzed missing data patterns and implemented various strategies such as **imputation** and **removal** to clean the dataset. Finally, we gained expertise in evaluating and selecting appropriate models, including **tree-based models**, **ensemble methods**, and **linear models**, to best suit the problem at hand.

7. Team Member Contribution

- Nipun Kothari : Data Preprocessing, Feature Engineering
- Nishant Singh : Exploratory Data Analysis, Deployment
- Rahul Bharti : Literature Review, Hyperparamter Tuning
- Ravada Satyadev : Baseline Model Experiment, Results and Conclusion
- Udbhav Singh : Feature Engineering, Hyperparameter Tuning

References

1. Al Mamlook, Rabia Kwayu, Keneth Frefer, Abdulbaset. (2019). Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity. 10.1109/JEEIT.2019.8717393. .

2. Assi, Khaled Rahman, Syed Masiur Mansoor, Umer Ratrout, Nedal. (2020). Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol. International Journal of Environmental Research and Public Health. 17. 10.3390/ijerph17155497.

3. Çelik, Ali Sevli, Onur. (2022). Predicting Traffic Accident Severity Using Machine Learning Techniques.