Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/52/20241582

Severity's prediction of car accidents in PA and model comparison

Wenchuan Chu[1,3,5], Xuanrun Qu[2,4]

[1]Lehigh University, 731 Evans Street, Bethlehem, PA, 18015
[2]University of California San Diego, 3151 via Alicante, La Jolla, 92037

[3]wcchu_scott@outlook.com
[4]chrisqu0316@gmail.com
[5]corresponding author

Abstract. This study aims to provide an analytical and predictive framework for understanding and forecasting the probabilities of varying severities of traffic accidents in Pennsylvania, based on the US Accident dataset dated up to March 23. Initially, the dataset and variable descriptions are provided for comprehensive understanding. During the data preprocessing phase, variables undergo a thorough examination for missing values, and suitable imputation methods are selected for data completeness. Further, feature selection and data cleansing are executed to prepare the dataset for model training. The study predominantly utilizes three machine learning algorithms, Logistic Regression, Random Forest, and Support Vector Machine, to construct predictive models. The performance of these models is meticulously evaluated for accuracy and compared through specific data sampling techniques. Overfitting checks, feature importance elucidation, and in-depth discussions on model performance variations are also included. By navigating through this multifaceted analysis, the study aims to shed light on the strengths and weaknesses of different modeling approaches for this particular problem context, thereby providing valuable insights for future research endeavors.

Keywords: Traffic accidents, Machine learning, Severity prediction, Feature selection, Pennsylvania.

1. Introduction

Traffic accidents have been a huge concern for decades, leading to numerous fatalities and injuries worldwide. As transportation networks expand, road traffic dynamics become increasingly complex; thus, understanding the patterns, causes, and consequences of traffic accidents can provide valuable insights for policymakers and urban planners to enhance road safety.

Recent advances in data collection techniques have led to vast datasets detailing various aspects of traffic accidents. Machine learning and statistical modeling offer powerful tools to make sense of this data, providing a comprehensive view of the determinants of traffic accident severity and enabling more effective interventions.

The paper is structured as follows: Section 1 introduces this project. Section 2 focuses on data pre-processing, including details on data acquisition, cleaning, and additional data processing steps. Section 3 discusses methodology, including elaborating on feature selection techniques and model creation.

DOI: 10.54254/2755-2721/52/20241582

Section 4 is about results analysis, providing an in-depth exploration of the outcomes, significant findings, and implications. Section 5 is the conclusion, summarizing the main takeaways, emphasizing the study's main contributions, and pointing out future research directions.

Several studies have shown that different machine-learning techniques can offer more accurate models for this complex prediction task.[1] Factors like road conditions, weather, and driver behavior, among others, have been considered in creating more accurate and robust models. However, deciding which machine learning model can offer the best prediction accuracy while balancing complexity and computational cost has not been easy. The choice of model is especially crucial because the computational resources involved can sometimes account for a significant portion of the total operational expenses.

Machine learning models like Random Forest, Logistic Regression, and Support Vector Machine (SVM) have been previously used for predicting traffic accident severity[2]. Random Forest offers the advantage of high accuracy but is hard to interpret. Logistic Regression is easy to interpret but could better deal with non-linear relationships between variables. Conversely, SVM is known for its high accuracy in high-dimensional spaces but is rather time-consuming.

Existing studies also present divergent results concerning the effectiveness of these models. Some suggest that Random Forest outperforms other models, while others argue that simpler models like Logistic Regression could be equally effective when feature engineering is applied carefully[3]. This raises questions about the relative importance of model complexity versus feature selection in predicting traffic accident severity.

Therefore, this study aims to empirically compare the effectiveness of Random Forest, Logistic Regression, and SVM in predicting the severity of traffic accidents and understand the role of feature importance in the performance of these models. A better understanding of these aspects will offer a more informed choice of model, potentially saving both computational resources and human lives.

The dataset utilized for this study offers a comprehensive look into traffic accidents across 49 states in the United States. The dataset, collected from February 2016 to March 2023, is amassed from various trusted sources, including state transportation departments, law enforcement agencies, traffic cameras, and road-network traffic sensors. By employing multiple APIs for real-time traffic event data collection, the dataset comprises approximately 7.7 million accident records, rendering it highly representative of traffic conditions and associated risk factors in the United States.

2. Data processing

For this part, we aim to explore the dataset of road accidents and prepare it for further analysis, focusing on predicting the severity of road accidents. The initial dataset has 296620 entries and 45 attributes. Each entry represents a unique road accident event, and each attribute represents features related to the accident, such as geographical coordinates, time, weather conditions, and road conditions. The info() method was used to get an initial understanding of the dataset's structure.

2.1. Missing ratio computation

Data cleaning was essential for maintaining dataset quality. We used merging techniques to combine cleaned numeric and object data, ensuring no missing values existed. The Dataframe df_clean was generated using the pd. concat () function. It shows that the dataset contained no missing values and had 296620 entries across 46 attributes. Addressing the class imbalance in the Severity variable is important since an unequal representation might lead to biased predictions. Therefore, techniques such as oversampling the minority class were employed to help solve this issue[4].

A df query was built to find the columns containing missing values and then to sort them by the ratio of missing values. Doing so showed nine columns containing missing values, with Wind_Direction having the highest missing ratio at about 0.022, followed by columns like Weather_Condition and Weather_Timestamp. Initial exploratory data analysis was conducted to help understand the data distribution, outliers, and general trends. The following figure is the table of the variables' missing ratio[5]:

DOI: 10.54254/2755-2721/52/20241582

Table 1. Missing ratio for the dataset

Index	Column_name	Missing_count	Missing_ratio
11	Wind_Direction	6521	0.022
12	Weather_Condition	6024	0.020
10	Weather_Timestamp	5115	0.017
13	Sunrise_Sunset	1578	0.005
14	Civil_Twilight	1578	0.005
15	Nautical_Twilight	1578	0.005
16	Astronomical_Twilight	1578	0.005
9	Airport_Code	523	0.002
2	Street	408	0.001

2.2. Data format unification

During dataset preparation, we addressed missing values in features like 'Wind_Direction,' 'Weather_Condition,' and 'Weather_Timestamp'. Missing data, less than 2.2% of the dataset, was imputed using the most frequent value for categorical variables and the mean for numerical ones.

After successful imputation, we recast the data types of various features to more suitable forms, such as 'uint8' for 'Timezone' and 'float32' for numeric features like 'Start_Lat,' to optimize memory usage and computational efficiency. Subsequently, certain features were transformed to improve their utility in the upcoming analyses. For instance, time-related columns like 'Sunrise_Sunset' were mapped to numerical representations ('Day': 1, 'Night': 0). We extracted detailed temporal elements like 'Year,' 'Month,' 'Day,' 'Hour,' and 'Minute' from the 'Start_Time' timestamp. We also calculated a new feature, 'Time_Duration(min),' by rounding off the time difference between 'Start_Time' and 'End_Time.'

However, it was observed that some entries had negative time duration values, which were clearly outliers. We addressed this by setting these entries to NaN and subsequently dropping them from the dataset. Finally, the cleaned numerical and categorical data frames were concatenated to form a comprehensive, cleaned dataset, ready for further analytical procedures.

By executing these preprocessing steps, we aimed to maximize the reliability and interpretability of the dataset, thereby providing a robust foundation for the next stages of our data science pipeline.

2.3. Additional Data Processing

Custom mapping based on frequency counts and Label Encoding were utilized for encoding. For columns like 'County,' 'Wind_Direction,' 'Month,' 'Weekday,' and 'Weather_Condition,' a two-step process was followed. First, the unique values within each column were sorted based on their occurrence frequency, from the least frequent to the most frequent. A custom mapping was created to replace each unique value with its corresponding index in the sorted list. For instance, the least frequent value would be mapped to 0, the next least frequent to 1, assigning a weightage based on the frequency of each category under the assumption that rarer categories might have more significance.

Then the sklearn's LabelEncoder was used to transform these columns. While the initial custom mapping was based on frequency, LabelEncoder ensured that the variables were encoded to values ranging from 0 to n_classes-1, where n_classes is the number of unique values. For columns of Boolean data types, these were straightforwardly converted to integer types: 0 for False and 1 for True. Finally, for the remaining columns with 'category' and 'string' data types, LabelEncoder was applied.

In summary, the encoding process aimed to convert all non-numeric variables into a numerical format while attempting to preserve as much semantic meaning of the data as possible. This step is crucial for ensuring the data is appropriately formatted for machine learning algorithms requiring numerical input features.

DOI: 10.54254/2755-2721/52/20241582

Table 2. Variables after manual selection

Column_name	Non-Null	Data Type
Severity	296620	uint8
Distance(mi)	296620	float32
Temperature(F)	296620	float32
Humidity(%)	296620	float32
Pressure(in)	296620	float32
Visibility(mi)	296620	float32
Wind_Speed(mph)	296620	float32
Amenity	296620	bool
Bump	296620	bool
Crossing	296620	bool
Give_Way	296620	bool
Junction	296620	bool
No_Exit	296620	bool
Railway	296620	bool
Roundabout	296620	bool
Station	296620	bool
Stop	296620	bool
Traffic_Calming	296620	bool
Traffic_Signal	296620	bool
County	296620	category
Wind_Direction	296620	category
Weather_Condition	296620	string
Sunrise_Sunset	296620	bool
Month	296620	category
Hour	296620	uint8
Weekday	296620	category
Time_Duration(min)	296620	float32

3. Methodology

3.1. Additional Data Processing

This section analyzes how to filter valid data. This paper mainly describes three methods: manual selection, VIF selection and ANOVA selection

3.1.1. Manual Process

The purpose was to ensure that only the most relevant features would be retained in the dataset, aiding in model interpretability and performance. To this end, we undertook a manual process to drop columns deemed less pertinent for the upcoming analytical tasks. Columns such as 'Description' and 'Weather_Timestamp' were removed primarily because they are high-dimensional text fields, which could introduce noise rather than meaningful variance. The 'Source' column was dropped since it was not considered to have predictive power for the problem at hand.

Additionally, location-based features like 'Street' and 'City' were removed due to their extensive cardinality, making them challenging to analyze and incorporate into predictive models effectively. The same rationale was applied for dropping 'State' and 'Zipcode.' 'Airport_Code' was eliminated as it was deemed irrelevant to the analysis goals.

DOI: 10.54254/2755-2721/52/20241582

Temporal features like 'Start_Time' and 'End_Time' were also omitted, given that we had already extracted essential time components such as 'Month,' 'Hour,' and 'Weekday' in the preprocessing stage. Columns containing latitude and longitude information ('Start_Lat', 'Start_Lng') were removed due to their collinearity with other geographical features. Similarly, other features such as 'Timezone,' 'Day' 'Min' and 'Year' were also eliminated for collinearity or limited predictive power. Lastly, 'Astronomical_Twilight,' 'Nautical_Twilight,' and 'Civil_Twilight' were deemed redundant in the presence of the 'Sunrise_Sunset' variable, which encapsulates similar information straightforwardly.

By rigorously selecting variables, we aimed to create a streamlined dataset that retains only the most impactful features, facilitating more efficient and insightful subsequent analyses.

3.1.2. VIF selection

Variance Inflation Factor (VIF) is a statistical measure used to quantify how much the variance of an estimated regression coefficient increases when your predictors are correlated[6]. In simpler terms, VIF gauges the extent to which the presence of a variable can be explained by other variables in the model. It is calculated as $VIF = 1 / (1 - R^2)$, where R^2 is the coefficient of determination for the regression of a given variable against all other variables. Generally, a VIF value greater than 10 indicates a problematic level of multicollinearity.

Applying VIF in this project proved highly effective for identifying collinearity among predictors. By leveraging VIF, we could focus on the most relevant predictors, avoiding overfitting and improving the model's performance. The VIF analysis allowed us to surgically remove variables that could compromise the interpretability and reliability of our model, such as 'Pressure(in)' and 'Weather_Condition'. Consequently, our model can now offer more precise and interpretable results, underscoring the utility of VIF as an instrumental feature selection tool in predictive modeling.

In our efforts to build a robust predictive model, variable selection is crucial to mitigate the risk of multicollinearity, thereby enhancing the model's explanatory power and prediction accuracy. For this, we employed the Variance Inflation Factor (VIF) as the metric to gauge the level of collinearity among the predictors. A common threshold for VIF is 10; values exceeding this possess high multicollinearity. Our analysis revealed several variables with exceedingly high VIF scores: 'Pressure(in)' with a VIF of 400.81, 'Weather_Condition' at 278.78, 'County' at 24.16, 'Humidity(%)' at 19.93, 'Severity' at 19.42, 'Visibility(mi)' at 17.76, and 'Temperature(F)' at 14.53. These variables were removed from the model to prevent inflated standard errors that could impair the model's interpretability and predictive power. In contrast, variables such as 'Distance(mi)', 'Wind_Speed(mph)', and 'Time_Duration(min)' demonstrated low VIF scores below 10, indicating that they are less prone to multicollinearity and thus were retained in the model.

The following figure is the result after VIF selection:

Table 3. Variable after VIF selection

#	Column_name	Non-Null	Data Type
0	Severity	296620	uint8
1	Distance(mi)	296620	float32
2	Wind_Speed(mph)	296620	float32
3	Amenity	296620	int64
4	Bump	296620	int64
5	Crossing	296620	int64
6	Give_Way	296620	int64
7	Junction	296620	int64
8	No_Exit	296620	int64
9	Railway	296620	int64
10	Roundabout	296620	int64

DOI: 10.54254/2755-2721/52/20241582

Table 3. (continued).

11	Station	296620	int64
12	Stop	296620	int64
13	Traffic_Calming	296620	int64
14	Traffic_Signal	296620	int64
15	Wind_Direction	296620	int64
16	Sunrise_Sunset	296620	int64
17	Month	296620	int64
18	Hour	296620	uint8
19	Weekday	296620	int64
20	Time_Duration(min)	296620	float32

3.1.3. ANOVA selection

ANOVA is a statistical method to test differences between means, and it was used here to test the differences between the means of the dependent variable, Severity, for all different categories of each independent variable[7]. F_classif is a function used here to compute the ANOVA F-value between features, so it is useful for feature selection while using SelectKBest. SelectKBest is a feature selection algorithm that can select the top k features with the most significant impact on the targets.

The variables from ANOVA selection are: 'Distance(mi)', 'Temperature(F)', 'Pressure(in)', 'Wind_Speed(mph)', 'Crossing', 'Junction', 'Stop', 'Traffic_Signal', 'County', 'Wind_Direction', 'Weather_Condition', 'Sunrise_Sunset', 'Month', 'Hour', 'Weekday'

3.2. Model Creation

A two-pronged modeling approach was used to understand the intricate dynamics underlying traffic accidents, leveraging Logistic Regression and Random Forest algorithms, offering distinct advantages; Logistic Regression is known for its interpretability and ease of implementation, while Random Forest brings higher flexibility and accuracy.

Two sets of features will be used to train each of these models:

1.VIF-Selected Features: The first set of features is carefully selected based on Variance Inflation Factor (VIF) to mitigate the impact of multicollinearity. This approach aims to enhance the robustness and interpretability of the model.

2.Manually-Selected Features: The second set comprises features that have been retained solely based on domain knowledge and pragmatic considerations.|

The core objective of employing these dual methodologies is to probe into the nuanced ways in which feature selection affects model performance and interpretability, which is particularly pertinent given that traffic accidents result from a multitude of interacting variables. Simplifying or altering the data can inadvertently affect the model's explanatory power.

Therefore, the overarching goal of this study is to evaluate and compare the performance and interpretability of Logistic Regression and Random Forest models when trained on distinct feature sets. Performance will be assessed through key metrics such as accuracy, precision, and recall, while interpretability will be gauged by examining feature importance rankings. The study aims to offer actionable insights for more effective traffic safety interventions through this integrated approach.

By examining model performance across different feature combinations, the study offers a comprehensive and in-depth perspective into the complexities of traffic accidents. This multi-model, multi-feature set approach delivers multiple angles of performance evaluation and enriches our understanding of which variables hold significant weight in the real-world applicability of the models.

DOI: 10.54254/2755-2721/52/20241582

3.3. Random Forest Model

Model Creation Random Forest is an ensemble learning method that constructs multiple decision trees and combines them for more accurate and stable predictions[8]. The algorithm uses Bootstrap aggregating, where each tree is trained on a randomly selected subset of data with replacement. Furthermore, at each decision node, the best split is determined not from all features, but from a random subset of them. For classification tasks, the prediction is the class voted by the majority of the trees; for regression, it's the average of all trees' predictions. Beyond its predictive power, Random Forest provides feature importance scores, doesn't require feature normalization, and can efficiently handle large datasets with numerous features. The formula of Random Forest is shown below:

$$RFclassification(x) = majority\{T1(x), T2(x), \ldots, Tn(x)\}$$

$$RFregression(x) = \frac{1}{n}\sum_{i=1}^{n} Ti(x)$$

Post-training, we evaluate the model's performance using key metrics such as accuracy. Feature importance is also extracted to determine which variables most significantly impact accident severity, thus shedding light on areas that may warrant further investigation or intervention. This approach not only provides us with a robust predictive model but also offers interpretable insights that can be invaluable for policy formulation in the area of traffic safety. Through this model, we aim to achieve high predictive accuracy while maintaining computational efficiency, thereby making it both scalable and practical for real-world applications.

4. Results Analysis

To fine-tune the model and extract the best hyperparameters, we employ Randomized Search Cross-Validation. A balance between computational efficiency and model performance guides hyperparameters choice for Randomized Search. We explore n_estimators (number of trees in the forest) between [50, 100] and max_depth (maximum depth of each tree) among [None, 10, 20] to facilitate model complexity. min_samples_split and min_samples_leaf are constrained to small integers [2, 5] and [1, 2], to prevent overfitting. The max_features parameter is set to either 'auto' or 'sqrt' to define the number of features to consider when looking for the best split[9]. Randomized Search is executed with 5 iterations and 2-fold cross-validation to balance runtime and reliability.

4.1. Feature Importance Analysis

The chart illustrates the top 10 features' importance in a Random Forest model, highlighting that "Time_Duration(min)" stands out as the most significant feature, followed closely by "Distance(mi)" and "County", while factors like wind direction and weather conditions exhibit lesser importance:

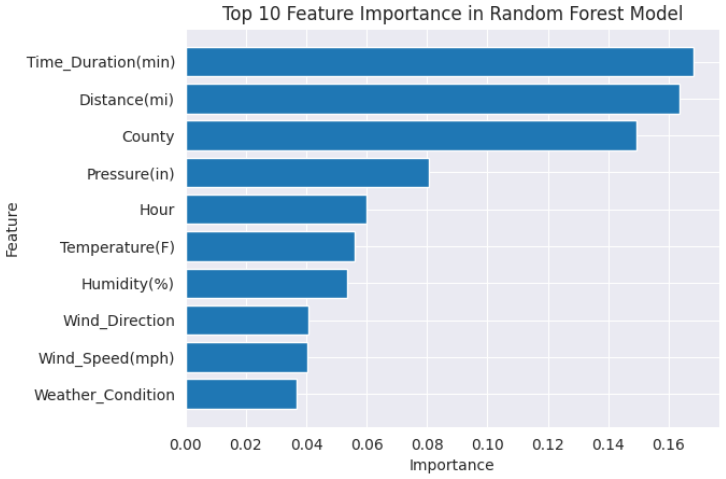

Figure 1. Feature Importance in Random Forest Model with manual selection's variables

DOI: 10.54254/2755-2721/52/20241582

The model's accuracy of 0.8799642640415346 (or approximately 88%) indicates that, out of all predictions made by the Random Forest model, about 88% of them were correct. This suggests a high degree of predictive reliability, demonstrating that the model can accurately classify or predict outcomes for the vast majority of the input data it encounters.

The presented graph depicts the top 10 features' importance in a Random Forest model, post Variable Inflation Factor (VIF) selection. It's evident that "Time_Duration(min)" takes precedence as the foremost influential feature. "Distance(mi)" and "Hour" also have notable importance, while attributes like "Junction" and "Stop" have less impact:

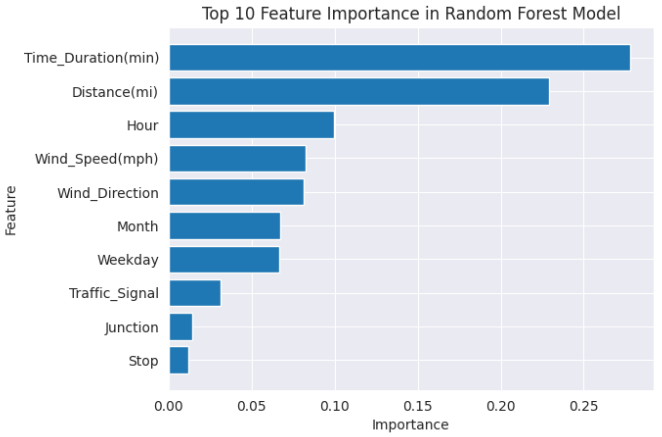

Figure 2. Feature Importance in Random Forest Model with VIF selection's variables

Additionally, the model boasts an accuracy of approximately 85.74% signifying its commendable predictive accuracy on the dataset after the VIF-based feature selection.

The dataset's severity distribution highlights that the majority of accidents fall under Severity Level 2, comprising a significant 83.61% of all incidents. While the least severe incidents (Severity Level 1) are relatively rare, accounting for just 0.58%, the most severe accidents (Severity Level 4) make up 5.35%. This distribution emphasizes the prevalence of moderately severe accidents and the imperative need for improved safety measures across all severity levels to mitigate risks:

Table 4. Unique value distribution in 'Severity' column

Severity	Unique Number	Proportion
1	23.56	0.583575
2	34.64	83.605623
3	23.76	10.461533
4	27.9	5.349268

4.2. Report Comparison

In the post-ANOVA selection analysis, the average predicted probabilities for different severity levels are presented. Severity Level 2 has the highest average predicted probability at 0.9033, indicating that most accidents are anticipated to fall into this category. This is corroborated by the histogram for Severity 2, which shows a pronounced peak nearing a probability of 1. Meanwhile, Severity Levels 1, 3, and 4 have lower predicted probabilities of 0.3347, 0.5379, and 0.3500 respectively. The variances in these probabilities suggest there's some level of uncertainty in predictions across the board, but particularly for Severity Level 1 with a variance of 0.0392. Visually, the histograms for Severity Levels 1, 3, and 4 demonstrate diverse distributions, with no single pronounced peak akin to Severity 2, indicating a more spread-out range of predicted probabilities for these categories:

DOI: 10.54254/2755-2721/52/20241582

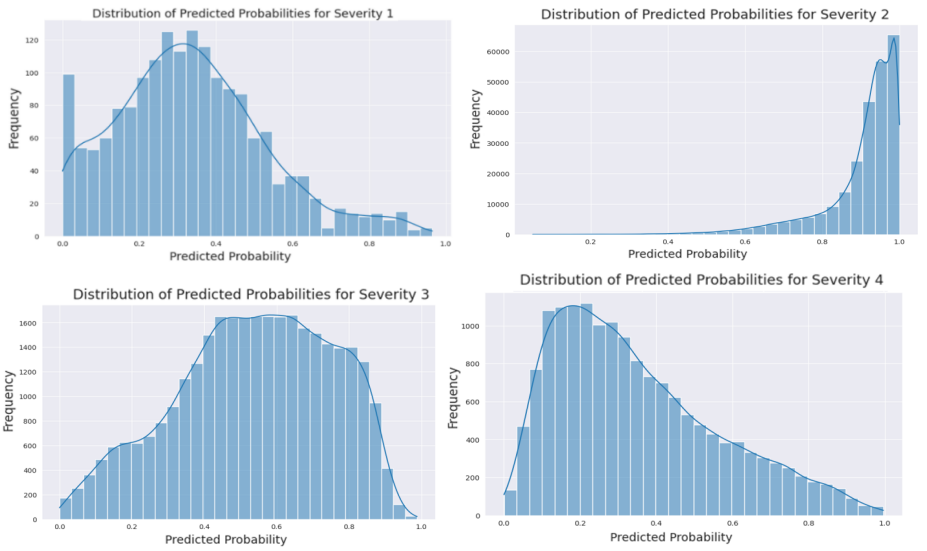

Figure 2. Predicted probabilities distribution for each 4 severity with ANOVA selection

Table 5. Result's statistic conclusion on ANOVA selection

Severity	Average predicted probability	Variance of predicted probability
1	0.3347	0.0392
2	0.9033	0.0106
3	0.5379	0.0458
4	0.3500	0.0466

Table 6. Result's statistic conclusion on Manual selection

Severity	Average predicted probability	Variance of predicted probability
1	0.3491	0.0372
2	0.9085	0.0104
3	0.5734	0.0424
4	0.3896	0.0527

The resultant data presents distinct predicted probabilities for the four severity levels. Severity Level 2 continues to dominate with a pronounced average predicted probability of 0.8905, supported by its histogram, which prominently peaks near a probability of 1, reiterating the high likelihood of events falling under Severity Level 2. On the other hand, Severity Levels 1, 3, and 4 register notably lower average probabilities of 0.2148, 0.4567, and 0.3288, respectively. These values are reflected in their corresponding histograms that portray diverse distributions. Severity Level 1, for instance, manifests a left-skewed distribution, suggesting fewer incidents reaching higher predicted probabilities. In contrast, Severity Levels 3 and 4 present more balanced distributions. The variances provided underline a level of uncertainty across all severity levels. Overall, the VIF-based selection further underscores the recurring prominence of Severity Level 2 in these predictions:

DOI: 10.54254/2755-2721/52/20241582

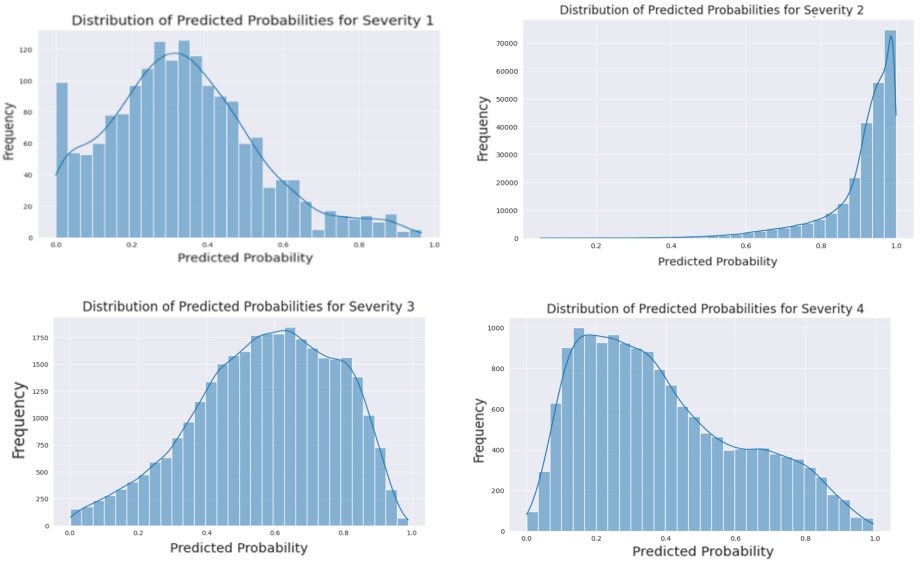

Figure 4. Predicted probabilities distribution for each 4 severity with Manual selection

Table 7. Result's statistic conclusion on VIF selection

Severity	Average predicted probability	Variance of predicted probability
1	0.2148	0.0256
2	0.8905	0.0110
3	0.4567	0.0417
4	0.3288	0.0464

Figure 5. Predicted probabilities distribution for each 4 severity with VIF selection

In our analysis of vehicle accident data in Pennsylvania from the U.S. dataset "US_Accident_March23," the dependent variable "Severity" revealed a notably imbalanced distribution.

DOI: 10.54254/2755-2721/52/20241582

Specifically, Severity level 2 accidents were highly over-represented, accounting for approximately 83.61% of the dataset, while Severity level 1 was significantly under-represented, constituting merely 0.58%. This imbalance might naturally lead to a higher predictive accuracy for Severity level 2 events in machine learning models, while diminishing the model's ability to accurately predict other Severity levels.

5. Conclusion and discussion

The significance of this dataset is more than providing data; it touches upon the public safety issue that accounts for a huge number of injuries every year. Given the dataset's wide-ranging geographic coverage and the many years of data collection, it serves as a strong foundation for examining the severity of traffic accidents. Moreover, the extensiveness of the dataset enables an evaluation of contributing factors such as road conditions, weather, and driver behavior, making the predictive models developed in this study[10].

Analyzing this huge dataset has academic relevance and offers benefits to society. Accurate models based on such data can provide invaluable decision-making support to government agencies, emergency services, and policymakers, potentially reducing traffic accidents' severity and frequency.

Upon model evaluation, several trends were evident. For example, with the ANOVA selection, the model offered an average predicted probability of 0.3347 for Severity level 1 events, with a variance of 0.0392. Since Severity level 1 incidents comprise only 0.58% of the actual dataset, the model's predictive performance is far from ideal in this context. For Severity level 2, the model's average predicted probability was robust at 0.9033, with a low variance of 0.0106. This resonates well with the dataset's high prevalence of Severity 2 events, confirming the model's adeptness in predicting this particular level. However, the average predicted probabilities for Severity levels 3 and 4 were 0.5379 and 0.3500, with variances of 0.0458 and 0.0466, respectively. These figures indicate a certain level of imprecision, especially in predicting Severity Level 3 events.

The observed imbalance in the distribution of the Severity variable has profound implications for predictive modeling tasks. Given the over-representation of Severity level 2 accidents, machine learning models are naturally predisposed to achieve higher predictive accuracies for this category. However, this bias comes at the cost of potentially undermining the model's competence in predicting rarer Severity levels, such as Severity level 1. This poses a challenge since predicting less frequent yet possibly more severe accidents is crucial for implementing effective preventive measures.

Understanding the inherent biases in the dataset is pivotal. It informs us about the data at hand and lays the foundation for more informed and strategic decision-making in model selection and training. Several strategies could be explored to enhance the robustness and fairness of our predictive models, in the future:

1. Resampling Techniques: We can consider oversampling under-represented severity levels or undersampling over-represented ones to achieve a more balanced dataset.

2. Cost-sensitive Learning: Models can be more attentive by assigning higher misclassification costs to under-represented severity levels.

3. Advanced Model Architectures: Exploring complex models or ensemble techniques that can handle class imbalances more effectively.

4. Incorporating Domain Knowledge: Integrating expert opinions or other external information can provide valuable context and enhance prediction accuracy for under-represented classes.

In conclusion, while the present dataset offers valuable insights into vehicle accidents in Pennsylvania, it also underscores the significance of addressing data imbalances for better predictive modeling. The journey of refining and optimizing our models is continuous, and addressing this challenge head-on will be pivotal in our endeavor to create safer roadways.

References

[1] Mair, Carolyn, et al. "An investigation of machine learning based prediction systems." Journal of systems and software 53.1 (2000): 23-29.

DOI: 10.54254/2755-2721/52/20241582

[2] Pradhan, Biswajeet, et al. "Modeling traffic accident severity using neural networks and support vector machines." Laser Scanning Systems in Highway and Safety Assessment: Analysis of Highway Geometry and Safety Using LiDAR (2020): 111-117.
[3] Strobl, Carolin, James Malley, and Gerhard Tutz. "An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests." Psychological methods 14.4 (2009): 323.
[4] Dhepe, Y. (2023). Project-4 US Accidents Data EDA. GitHub. Retrieved September 20, 2023, from https://www.kaggle.com/code/yuvrajdhepe/project-4-us-accidents-data-eda/notebook
[5] Moosavi, S. (2023). US Accidents (2016 - 2023): A Countrywide Traffic Accident Dataset (2016 - 2023). Kaggle. Retrieved from https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
[6] Oke, J., W. B. Akinkunmi, and S. O. Etebefia. "Use of correlation, tolerance and variance inflation factor for multicollinearity test." GSJ 7.5 (2019).
[7] Kim, Tae Kyun. "Understanding one-way ANOVA using conceptual figures." Korean journal of anesthesiology 70.1 (2017): 22-26.
[8] Seni, Giovanni, and John Elder. Ensemble methods in data mining: improving accuracy through combining predictions. Morgan & Claypool Publishers, 2010.
[9] Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A Countrywide Traffic Accident Dataset. arXiv preprint arXiv:1906.05409. Retrieved September 20, 2023, from https://doi.org/10.48550/arXiv.1906.05409
[10] Ebrahimi-Khusfi, Zohre, Ali Reza Nafarzadegan, and Fatemeh Dargahian. "Predicting the number of dusty days around the desert wetlands in southeastern Iran using feature selection and machine learning techniques." Ecological Indicators 125 (2021): 107499.