Crash Insights: Analyzing Car Accidents and Predicting Accident Severity in the US

Team 11: Deepanshu Jain, Sujith Kamme, Derek Lim and Timothy Sujo

Problem Definition

Car accidents remain a pressing concern in today's urban transportation environment, impacting both public safety and road conditions. To enhance road safety and develop proactive measures, it is imperative to examine the causal variables of accidents and understand their gravity. Our project aims to leverage extensive data to explore the nuances of car accidents, unveiling patterns, critical factors, and constructing a predictive model for accident severity. Our report addresses the following questions:

1. What are the top 10 most influential factors for accidents?
2. Are there certain times of the day, week, or month that have a higher frequency of accidents?
3. Are there any specific weather patterns that may lead to an increase in severity?
4. Are there any specific road features that lead to accidents?
5. What recommendations or interventions can be proposed to reduce the occurrence of accidents?

Background Description

This "Crash Insights" project is a response to this challenge, focusing on a meticulous analysis of the US Accidents Dataset [1][2]. The project's primary objectives are to uncover underlying patterns, identify critical contributing factors, and develop a predictive model for accident severity. According to a report [3], there were 39,508 fatal motor accidents in the United States in 2021 in which 42,939 people died. This amounted to 1.37 deaths per 100 million miles travelled and 12.9 deaths per 10,000 people.

Dataset Description:

Quantity and Quality - US Accidents Dataset is a comprehensive countrywide collection of data on auto accidents in the US from 2016 to 2021. It can be downloaded from here. Several APIs that continuously feed real-time traffic incident data were used to construct this dataset. Its large size makes it a perfect resource for assessing and forecasting the severity of accidents in the US.

License: CC BY-NC-SA 4.0 Dataset Size: 1.1 GB (2.8 Million Records)

Coverage: 49 States in the US Period: Feb 2016 to Dec 2021

Important Attributes - Key attributes include accident severity (target variable), geographical coordinates, weather conditions, and time-related features.

Experiment Setup and Analysis Results

1. *Data Identification*: We explored all the features along with their respective meanings and recognized the target variable as Severity, taking the following ordinal values: 1, 2, 3, and 4 (1 = Least Severity, 4 = Most Severity). We also dropped some columns like ID, Description, etc. irrelevant to our analysis.

2. *Univariate Analysis*: For numerical data, we utilized key statistics like mean, standard deviation, and box plots, etc, which helped us grasp the data's distribution and detect potential outliers in the features Pressure, Wind_Speed, Precipitation, and Visibility. Categorical variables with fewer than 25 categories were visualized through bar charts. Our analysis revealed several insights including the fact that "Calm" was the most common wind direction and Severity Level 2 was most prevalent throughout the dataset {shown in fig 1}.

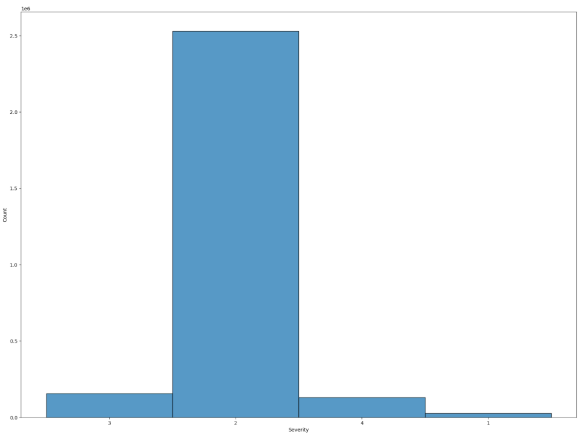

Fig. 1: Severity Distribution

For boolean variables, we tallied the occurrences of "true" and "false". Notably, "Crossing" had more accidents compared to "Railway". We also removed the "Turning_Loop" feature, which had a single value, "False."

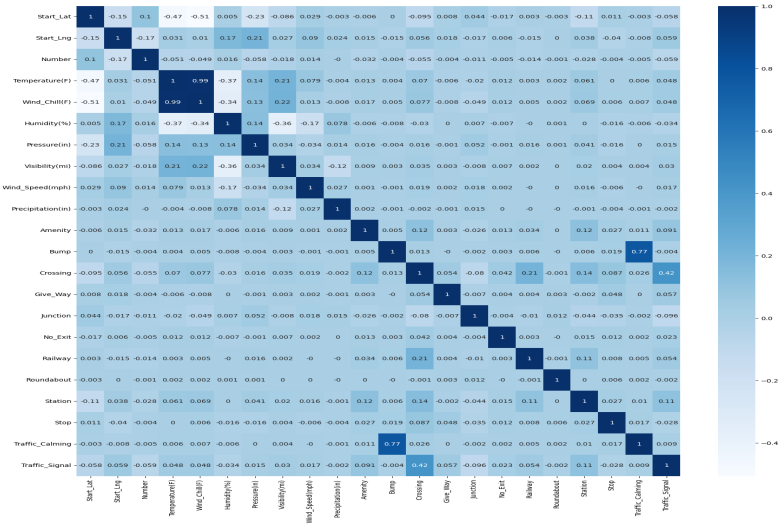

Fig 2: Correlation Heatmap

3. *Bivariate Analysis:* We explored the interplay of numerical variables via a heatmap {shown in fig 2}, revealing substantial correlations between certain variables.

Furthermore, we examined categorical variables containing weather, and road features, in relation to accident severity, revealing associations between them, which are shown later below in the dashboard.

4. *Missing Values Treatment, Removing Duplicate Data, and Outliers Treatment*: There were missing values in many columns. One column, “Number” (the street number), had 60% missing values and was dropped since it was irrelevant to our analysis. For other columns, we determined that median imputation was most suitable for numerical features, while mode imputation was used for many categorical ones. Specifically, for the Street column we utilized reverse geocoding to impute missing values.

We identified and dropped 175,921 duplicate rows. We tackled outliers in the Pressure and Wind_Speed columns by applying the IQR method.

5. *Variable Transformation and Creation*: We converted columns with only 2 categories (“Side”, “Sunrise_Sunset", "Civil_Twilight", "Nautical_Twilight", and "Astronomical_Twilight") to categorical variables using the dummy variables method using Pandas. To address the pronounced skewness in the

Precipitation and Visibility distributions, we applied logarithmic transformations, resulting in a significantly reduced skew and improved suitability for bivariate scatter analysis alongside other variables {shown in the fig 3}.

We also separated the Start_Time column into separate Year, Month, Weekday, and Hour columns for further time series analysis.

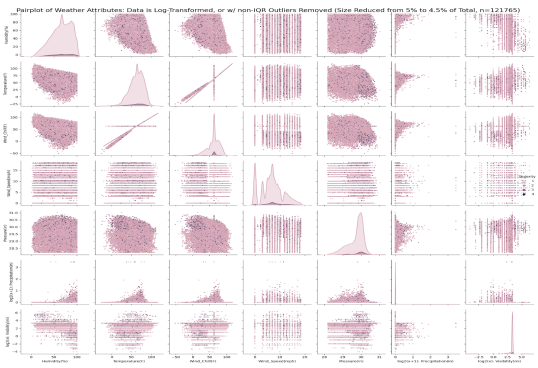

Fig 3: Pairplot after Log Transformation

6. *Exploratory Data Analysis and Visualization*:

I. Location Analysis:

We utilized several libraries such as Folium, and Geoplot to make such visualizations, and then observed that California had the most accidents followed by Florida and Texas {shown in fig 4}.

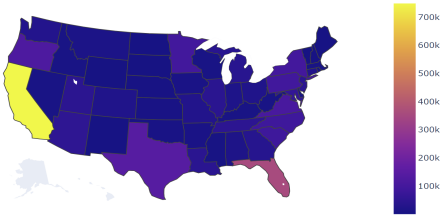

Fig 4: No. of Accidents State Wise

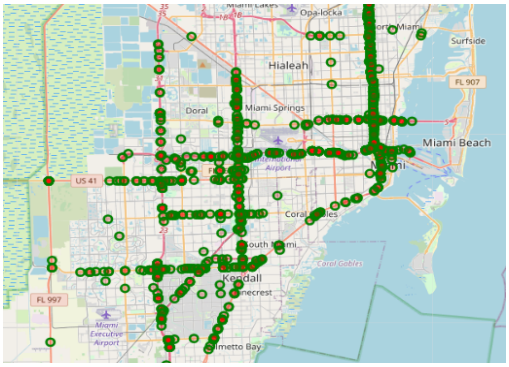

Fig 5: Accidents in Miami City

Based on cities, Miami city recorded the highest number of accidents followed by Los Angeles and Orlando.

To dig deeper into Miami to understand patterns in accidents, we found out that there are significantly higher numbers of accidents that occur on freeways {shown in fig 5}.

II. Time Series Analysis:

We observed that the number of accidents is growing at approximately 31% averagely per year. December has seen more accidents followed by November and January. Weekdays experience significantly more accidents compared to weekends, with weekend accident frequencies being at least 2/3 times lower [3]. The majority of accidents took place between hours 12 and 18{shown in Fig. 6}. This could be due to the increased traffic on roads and freeways especially during weekdays, as people have to fight rush hour traffic on the way back from work.

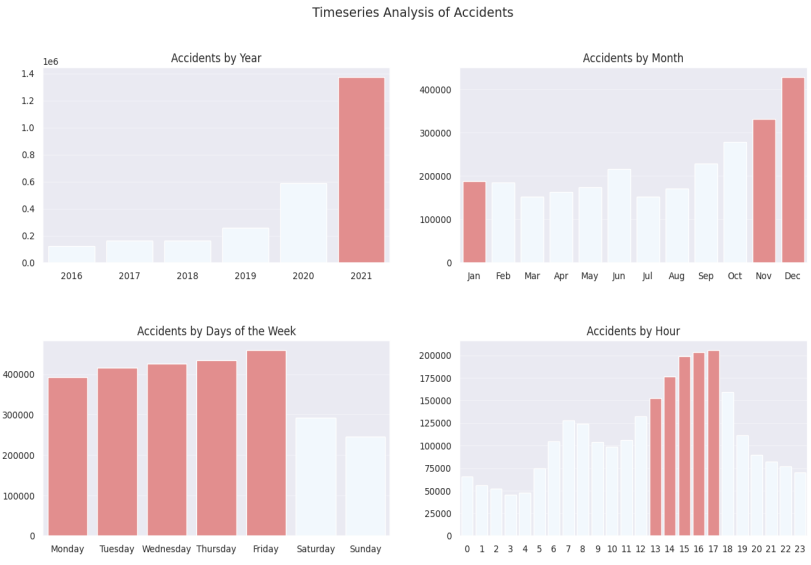

Fig 6: Time Series Analysis

III. Analysis on Weather Conditions and Road Features:

We analyzed the impact of wind speed, precipitation, and visibility {shown in Fig. 7} and found that although there are outliers, the main information here is that the severity of accidents shows, more often than not, a very low correlation with these driving conditions.

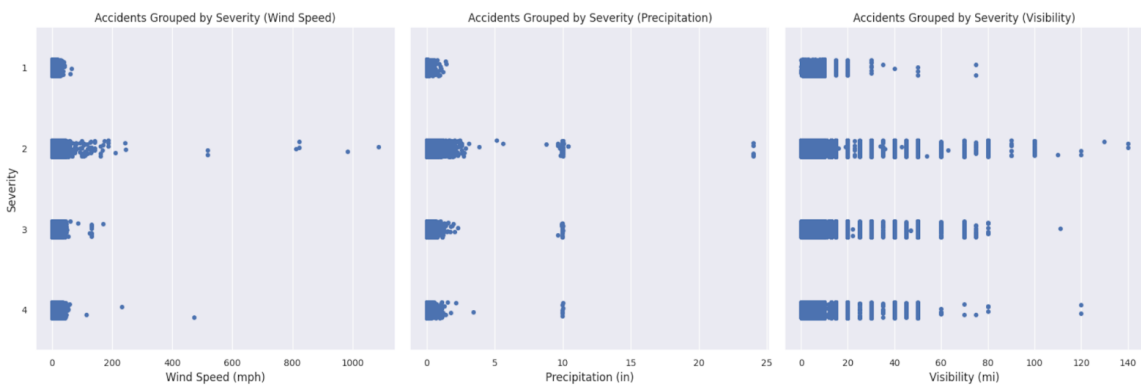

Fig 7: Impact of Wind Speed, Precipitation and Visibility on Accidents

Furthermore, we explored the breakdown of weather conditions by accident severity, showing that the most common weather conditions like "Clear" and "Cloudy" have more accidents overall by far. However, after normalizing the counts, further analysis revealed that the "Hail" and "Snow" weather conditions contained the highest proportions of Level 3 and 4 Severity compared to other weather conditions. With regards to road features, we observed that road junctions recorded the most no. of accidents followed by accidents at traffic signals and crossing {shown in Fig 8}[4].

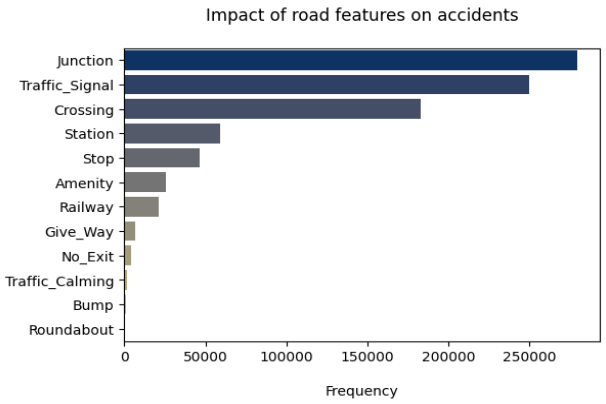

Fig 8: Impact of Road Features on Accidents

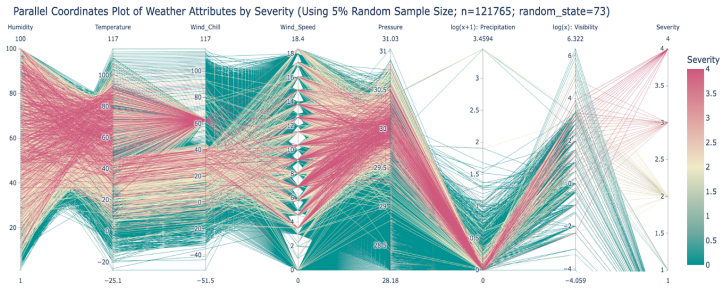

Fig 9: Parallel Coordinates Plot of Weather Attributes

IV. Multivariate Analysis of Numerical Weather Features: We also visualized trends of samples of the multidimensional numerical weather data {shown in fig 9}, showing how these columns are not so easily separable and cannot be used alone to classify accident severity.

V. What recommendations or interventions can be proposed based on the analysis to reduce the occurrence of severe accidents?

To reduce severe accidents, enhance high-traffic road infrastructure, implement advanced driver assistance systems, and engage communities with targeted awareness campaigns tailored to regional challenges like

adverse weather conditions (i.e. in cities with high hail or snow, talk about safe driving practices in these conditions)

7. *Feature Engineering*: We took a sample of the dataset for our model due to the huge size of the data, considering processing and storage constraints. Then, we split the dataset and performed feature encoding, ensuring that there was no data leakage. Next, we encoded the categorical variables for our model using LabelEncoder depending upon the number of attributes in each and considering its pros and cons.

Further, despite the use of a pipeline to balance instances of all classes to the number of instances of class Severity level 3, combining both the SMOTE techniques like RandomOverSampler and RandomUnderSampler, the expected improvement in results was not achieved. It posed a risk of overfitting by deviating from the true representation of the minority class, due to the generation of artificial instances. Consequently, we chose to proceed with the imbalanced dataset.

8. *Model Creation and Feature Standardization*: We decided to test multiple machine learning models, and see how each model performed when classifying accident severity. These models included KNN, Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Extreme Gradient Boosting (XGBoost). For models that required scaled features such as KNN, and Logistic Regression, we scaled the training features using StandardScaler. StandardScaler allows all our data to be standardized so that all variables contribute to the model evenly. For example, KNN relies on distance between data points, and if the features are on different scales the algorithm may give weight to figures with larger samples. For Logistic Regression, regularization may unfairly penalize certain features more than others. Consequently, standard scaling will normalize these features so that we have an efficient model.

Certain models didn't require any scaling, such as RandomForest, DecisionTrees, SupportVectorMachine, XGBoost, etc. For SVM, it is not required, but highly recommended for speed and performance improvement.

9. *Deep Learning Model:* Though the RandomForest and XGBoost models were performing well, we fitted the data to a multilayer perceptron model in PyTorch to see how the results would compare. We tested many different multilayer perceptron architectures that included different types of dropout layers, varied hidden dimensions, and activation functions, but ultimately the highest-performing deep learning model was a 5-layer multilayer perceptron using Leaky ReLu as the activation function.

10. *Model Evaluation and Improvement:* We found out that RandomForest and XGBoost turned out to be the best among the traditional ML models, while XGBoost had slightly higher scores compared to RandomForest. The neural network achieved an accuracy of 90% and an F1 score of 88%.

Though this neural network does well with classifying accident severity, it is slightly overkill for our dataset, as the model is extremely complex and time-consuming to code and train while only achieving similar results (and no improvement) to our high-performing classical machine learning models.

We achieved the following results which are shown in the above table{Table I}. But in this case, as we had imbalanced data, accuracy was not the best evaluation parameter for our model.

Top Models	*Accuracy*	*Precision*	*Recall*	*F1*
XGBoost	*0.90*	*0.87*	*0.90*	*0.87*
Random Forest	*0.90*	*0.87*	*0.90*	*0.86*
5 Layer Neural Net	*0.89*	*0.88*	*0.89*	*0.88*
Decision Tree	*0.84*	*0.85*	*0.84*	*0.84*

Table I: Model Evaluations

Based on our case, recall, and precision matter the most, therefore the F1-score is the best, assessing the overall ability to distinguish between classes and we have the majority of the cases as level 2 Severity.

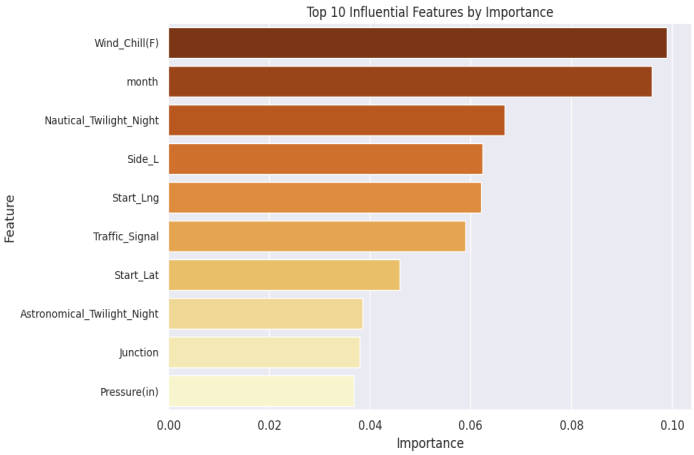

Fig 10: Influential Features for Model

We also concluded the most influential features {shown in Fig 10} for our XGBoost model consisted of columns like Wind_Chill, month, and Nautical_Twilight_Night, but all of them had low importance scores.

After this step, we tried improving our traditional models using hyperparameter tuning by GridSearchCV and even ensemble learning using SoftVotingClassifier and HardVotingClassifier.

11. ***An Interactive Dashboard*****:** An interactive dashboard was created using Tableau to provide a user-friendly interface for exploring the project's insights{shown in Fig 11}. It can be accessed from here.

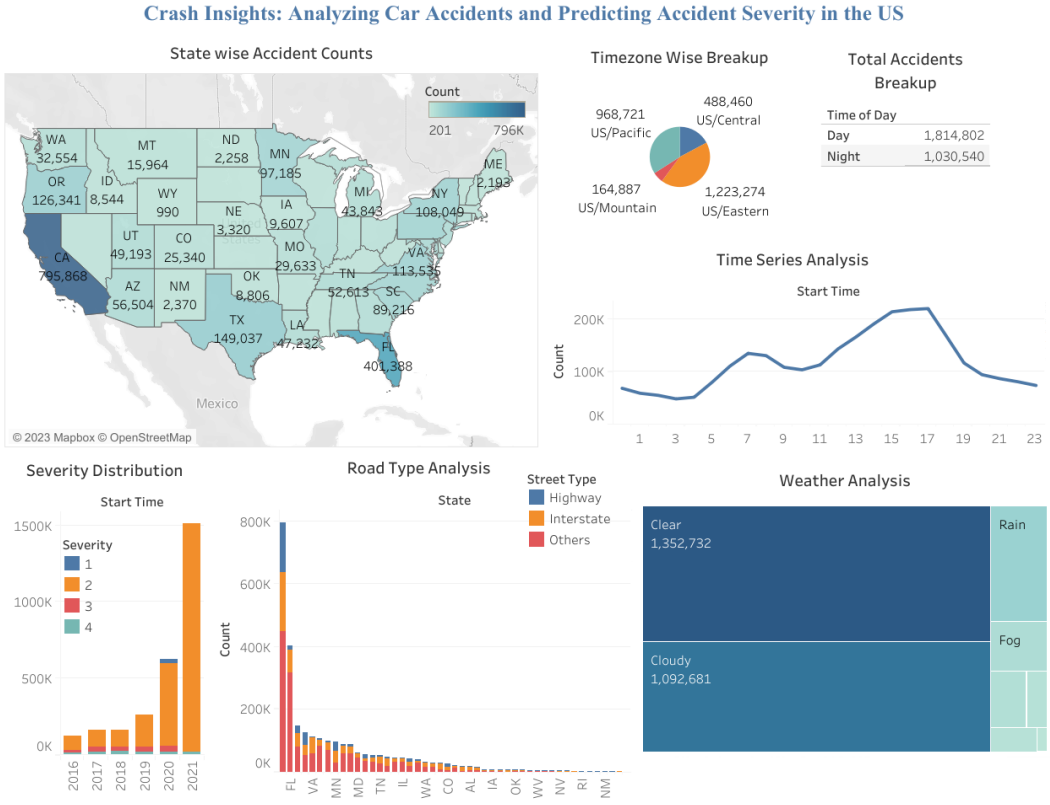

Fig 11: Interactive Dashboard

Technology Stack Used

Python, Numpy, Pandas, Matplotlib, Seaborn, Plotly, Folium, SweetViz, Tableau, Imblearn, Scikit-learn, Google Colab, XGBoost, PyTorch

Observation and Conclusion

Our project uncovered crucial patterns in accident data, highlighting key factors affecting severity. Location and time emerged as pivotal [5], with certain areas and specific temporal periods linked to higher severity. This temporal understanding could be useful for optimizing traffic management during peak risk times. We found how important features like road characteristics and weather conditions contribute to severity, emphasizing the need for weather-aware traffic systems and awareness campaigns for cautious driving in challenging weather. These insights enable a holistic understanding of accident dynamics, supporting a more informed approach to road safety.

Our project successfully developed an accurate predictive model through advanced analytics and machine learning, translating patterns into actionable insights. These types of models have far-reaching implications and could be used to guide local authorities in enhancing infrastructure, implementing targeted awareness campaigns, and optimizing emergency response strategies. Additionally, similar models could enable real-time risk assessment and intervention, contributing to the overarching goal of creating safer roadways.

Future Scope

In the future, we could work on using the text contained in the dataset's "Description_Provided" attribute to perform text segmentation for further data mining, analysis, and feature creation. This project can also be extended to incorporate real-time prediction capabilities, leveraging emerging technologies and continuous data streams. Continuous updates to the dataset and the adoption of more sophisticated machine learning models could further refine the accuracy and applicability of the predictive model. In the future, we can also utilize hierarchical clustering to classify the hot spots based on the roadways and surrounding lands.[5]

References

1. Moosavi, Sobhan, et al. "A Countrywide Traffic Accident Dataset." arXiv.Org, 12 June 2019, arxiv.org/abs/1906.05409.
2. Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019, September 19). *Accident risk prediction based on heterogeneous SPARSE DATA: NEW DATASET and insights*. arXiv.org. https://arxiv.org/abs/1909.09638
3. "Fatality Facts 2021: State by State." *IIHS*, www.iihs.org/topics/fatality-statistics/detail/state-by-state. Accessed 8 Dec. 2023.

4. Vozniuk, A., & Kaskiv, V. (2020, October 16). *Use of big data for actualization of approaches to road accident analysis*. SSRN. https://ssrn.com/abstract=3681353
5. Schneider, R. J., Sanders, R., Proulx, F., & Moayyed, H. (n.d.-a). *United States fatal pedestrian crash hot spot locations and characteristics*. Journal of Transport and Land Use. https://doi.org/10.5198/jtlu.2021.1825