Predicting Accident Severity: Report

Aayush Ranjan, Jyotirmaya Singh, Sanmay Sood, Siddharth Rajput
Indraprastha Institute of Information Technology, Delhi

1. Overview of the Dataset and Problem Statement

The dataset covers **1.1 million+ accidents** reported in 49 U.S. states from **February 2019 to March 2023**, sourced via **MapQuest API**. It includes 46 features categorized into Traffic, Address, Weather, POI, and Period-of-Day attributes. The target variable is **Severity (1 to 4)**, representing accident severity levels (with 1 being the least severe to 4 being the most severe).

Project Goals

- Identify factors influencing accident severity.
- Develop predictive models that forecast accident severity without relying on vehicle or driver details.

2. Challenges Faced & Solutions

- **Handling Missing Data:** Some weather features like *Temperature* and *Visibility* had missing values.
  Solution: Imputation based on *Airport_Code* and *Start_Month* using median values.
- **Dimensionality and Irrelevant Features:** Columns like *ID*, *End_Time*, and *Turning_Loop* offered little predictive value.
  Solution: Dropped irrelevant features and also used **L1 regularization** for feature selection in our comparative study.
- **NaN Handling:** Certain column data points were missing for instances of records. **Solution:** Rows with missing values in less significant features (like *City*, *Zipcode*, and various *twilight features*) were dropped.

3. Hypothesis Tests and Conclusions

Chi-Square Tests for Independence

The Chi-Square Test for Independence is a statistical test used to determine whether there is a significant association between two categorical variables. It assesses whether the observed frequency distribution of one variable differs across the levels of another variable.

- **Roundabout:** Null hypothesis (H_0): Severity and roundabouts are independent.
  Result: $p = 0.445 \rightarrow$ Fail to reject H_0.
  Explanation: Roundabouts have no significant impact on accident severity.
- **Heavy Rain:** Null hypothesis (H_0): Severity and heavy rain are independent.
  Result: $p \approx 0 \rightarrow$ Reject H_0.
  Explanation: Heavy rain shows a strong association with accident severity, likely due to poor visibility and slippery roads.
- **Fog:** Null hypothesis (H_0): Severity and fog are independent.
  Result: $p \approx 0 \rightarrow$ Reject H_0.
  Explanation: Foggy conditions significantly influence accident severity due to reduced visibility.

z-Test for Proportions

The z-Test for Two Proportions is a statistical test used to determine whether there is a significant difference between the proportions of a specific outcome in two independent groups. It compares the observed proportions to evaluate if the difference is due to chance or a real effect.

- **Heavy Snow:** Null hypothesis (H_0): No difference in severe accident proportions between heavy snow and non-heavy snow.
  Result: $p \approx 0 \rightarrow$ Reject H_0.
  Validation Experiment: Severe accident proportion is **39.76%** during heavy snow vs. **27.54%** otherwise.
  Explanation: Heavy snow increases the likelihood of severe accidents due to dangerous driving conditions.

t-Test for Visibility Between Mild and Severe Accidents

The t-Test for Two Population Means is a statistical test used to determine whether there is a significant difference between the means of two independent groups. It compares the sample means, accounting for variability and sample sizes, to assess if the observed difference is due to chance or a true effect.

- Null hypothesis (H_0): No significant difference in mean visibility between mild and severe accidents.
  Result: $p \approx 0 \rightarrow$ Reject H_0.
  Validation Experiment: Mean visibility for mild accidents = **9.53 miles**, severe accidents = **9.08 miles**.
  Explanation: Visibility is slightly worse for severe accidents, suggesting poor visibility contributes to severity.

t-Test for Correlation Between Temperature and Humidity

The t-Test for Correlation is a statistical test used to determine whether the correlation coefficient between two variables is significantly different from zero. It assesses the strength and direction of a linear relationship to evaluate if the observed correlation is likely due to chance or a true association.

- Null hypothesis (H_0): No correlation between temperature and humidity.
  Result: Correlation coefficient = **-0.35**, T-statistic = **-305**. There is a correlation between Humidity and Temperature.
  Explanation: A moderate negative correlation exists, where higher temperatures tend to lower humidity.

z-Test for Proportion of Nighttime Severe Accidents

The z-Test for Proportions is a statistical test used to determine whether the observed proportion of a specific outcome in a sample differs significantly from a hypothesized proportion or between two groups. It assesses if the difference is likely due to chance or reflects a true effect.

- Null hypothesis (H_0): The proportion of nighttime severe accidents is $\leq 60\%$.
  Result: Observed proportion = **36%**.
  Fail to reject the null hypothesis: Proportion of night-time severe accidents is not significantly greater than 0.6
  Explanation: Severe accidents are not any less frequent at night than at other times of the day. This could be due to lesser number of cars and congestion occurring at night but at the same time poor visibility may contribute to more severe accidents. During other times of the day, the mentioned factors have the opposite effect.

4. Model Performance: Accuracy and Time Comparison

Models Used

- **Random Forest:** An ensemble method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting.

- **Logistic Regression:** A linear model used for binary or multi-class classification that estimates the probability of a class based on input features.
- **Naive Bayes:** A probabilistic classifier based on Bayes' theorem, assuming independence between features. It works well for text classification and other categorical data.
- **K-Nearest Neighbors (KNN):** A non-parametric algorithm that classifies a point based on the majority class among its nearest neighbors.
- **XGBoost:** An optimized gradient boosting algorithm that builds decision trees sequentially to correct errors of previous iterations, ensuring high performance.
- **MLP (Multi-Layer Perceptron):** A neural network consisting of multiple layers with non-linear activation functions, capable of learning complex relationships in data.

Reduction Techniques

- **SVD (Singular Value Decomposition):** A linear dimensionality reduction technique that factorizes the data matrix into singular vectors and singular values. It retains the most important components (largest singular values) while reducing noise and computational load.
- **PCA (Principal Component Analysis):** A method that transforms data into orthogonal principal components ordered by the amount of variance they explain. PCA reduces dimensions while preserving as much data variability as possible.
- **JL (Johnson-Lindenstrauss Lemma):** A random projection-based method that reduces the dimensionality of the data while approximately preserving pairwise distances between points with high probability. It is fast and scalable for large datasets.
- **Feature Selection:** A technique that selects a subset of the most relevant features by removing irrelevant, redundant, or less informative features. Methods like L1 regularization (LASSO) can be used to shrink coefficients to zero.

Accuracy Comparison

Table 1: Model Accuracy Comparison

Model	**Original**	**SVD**	**PCA**	**JL**	**Feature**
Random Forest	81.31%	72.47%	73.33%	71.51%	76.61%
Logistic Regression	67.26%	60.65%	64.09%	61.62%	57.62%
Naive Bayes	52.06%	57.96%	51.34%	52.31%	58.58%
K-Nearest Neighbors	68.93%	68.48%	68.77%	68.54%	75.38%
XGBoost	75.05%	65.41%	68.11%	67.15%	72.09%
MLP	72.00%	63.54%	68.68%	67.77%	44.45%

Time Taken Comparison

Table 2: Model Training Time Comparison (in Seconds)

Model	**Original**	**SVD**	**PCA**	**JL**	**Feature**
Random Forest	76.86 s	19.49 s	41.32 s	37.12 s	7.26 s
Logistic Regression	41.98 s	1.00 s	7.54 s	5.49 s	33.14 s
Naive Bayes	0.56 s	0.09 s	0.26 s	0.23 s	0.06 s
K-Nearest Neighbors	70.05 s	15.43 s	46.77 s	49.04 s	3.89 s
XGBoost	4.70 s	2.68 s	4.47 s	4.80 s	2.14 s
MLP	32.61 s	32.07 s	33.91 s	31.61 s	32.88 s

Observations

- **Naive Bayes:** Fastest model but least accurate.
- **Random Forest:** Best accuracy, significant time improvement with scaling.
- **KNN:** Speed and accuracy improve with scaling.
- **XGBoost:** Balances speed and performance well.
- **MLP:** Stable training time, accuracy drops with scaling.

5. Conclusion & Future Work

Conclusion

- **Random Forest** demonstrated the highest accuracy.
- Weather factors like **Heavy Rain**, **Fog**, and **Heavy Snow** significantly impact severity.
- Scaling benefits simpler models like **KNN** but harms complex models like **MLP**.

Future Work

- **Feature Engineering:** Include traffic congestion and seasonal trends.
- **Model Enhancements:** Explore deep learning techniques.
- **Real-Time Deployment:** Optimize models for real-time predictions.

References

1. Dataset Source: MapQuest API.
2. Project Code Repository: GitHub Link.
3. Dataset Website: Link
4. Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019). *Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights.* In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 33–42). ACM. DOI: 10.1145/3347146.3359078.
5. Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). *A Countrywide Traffic Accident Dataset.* arXiv preprint. arXiv:1906.05409.