U.S. ACCIDENT SEVERITY ANALYSIS

Yichen Guo, Nianqing Chen, Zijing Cheng

INTRODUCTION

US-Accidents data offers a wealth of opportunities for applications such as real-time accident prediction, hotspot location analysis, casualty assessment, and rule extraction to predict accidents and study environmental influences on accident occurrence. By pinpointing high-risk areas, we can enhance driver awareness and caution in these regions. For instance, in a state with a high prevalence of speed-related accidents, drivers can be encouraged to adhere to speed limits and stay vigilant for potential speeders.

In this project, our primary focus is on identifying factors that influence accident **Severity**, which is classified on a scale of 1-4, with 1 representing the least impact on traffic (i.e., short traffic delays). This concise introduction sets the stage for a comprehensive academic poster presentation on our findings.

DATA

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to Dec 2021, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 2.0 million accident records in this dataset. The dataset is being subset due to computation time issues.

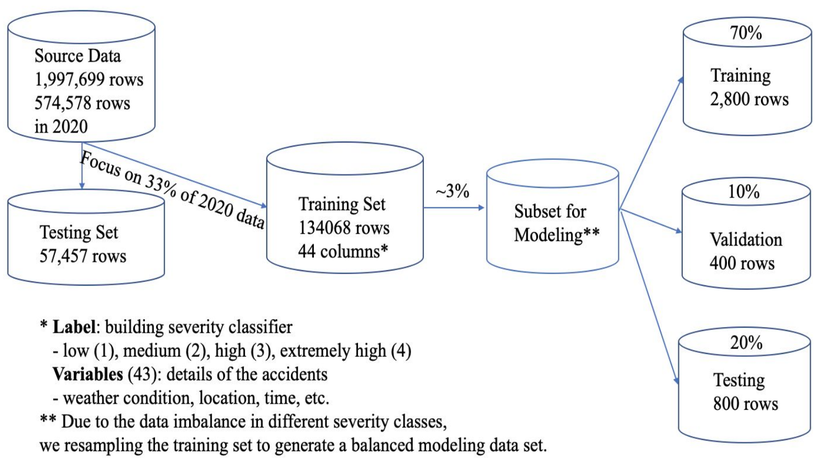

Figure 1: The Data Processing of the U.S. Accident dataset

METHODOLOGY I

Step 1: Using **cross validation** to select the top performer of the models then fine tuning the model to reach the best result in predicting the severity. Since the main research question is a multiclass classification problem, the method we are going to tried out are mostly classification method.

- Logistic Regression
- LDA
- QDA

- Bagging
- Naive Bayes
- Decision Trees

- Random Forest
- Ada Boosting
- Neural Networks

However after applying the models on the dataset, only 7 out of 9 models are applicable on predicting the multi-class classification problems. QDA and Ada Boosting need extra adjustment on their structure and will added in future analysis.

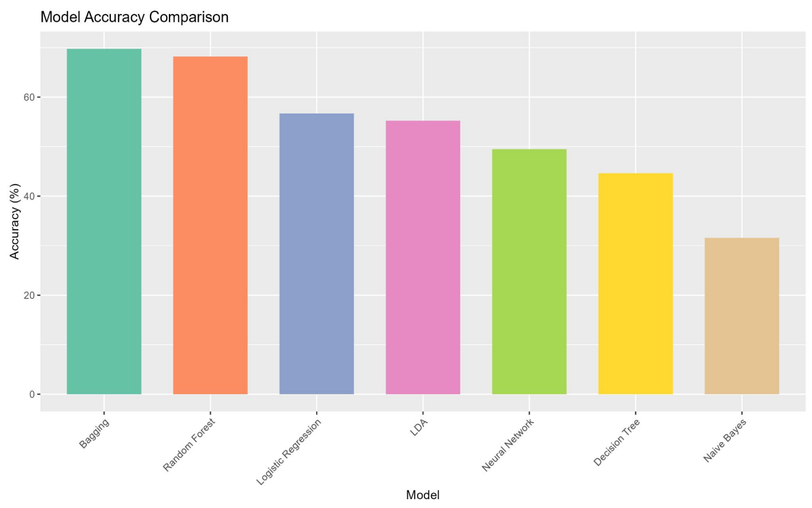

Figure 2: Accuracy comparison of each model using cross-validation

Model	AUC	Mean_Accuracy	AUC_Rank	Accuracy_Rank
Bagging	0.8886	0.7889	1	1
Random Forest	0.8827	0.7692	2	2
Logit Reg	0.7933	0.7113	3	3
LDA	0.7794	0.7016	4	4
Neutral Network	0.7498	0.6628	5	5
Decision Tree	0.6845	0.6299	7	6
Naive Bayes	0.7205	0.5441	6	7

Table 1: AUC and Accuracy comparison across models

METHODOLOGY II

Step 2: Train and tuned top 4 classifiers (Bagging, Random Forest, Logistic regression and LDA)

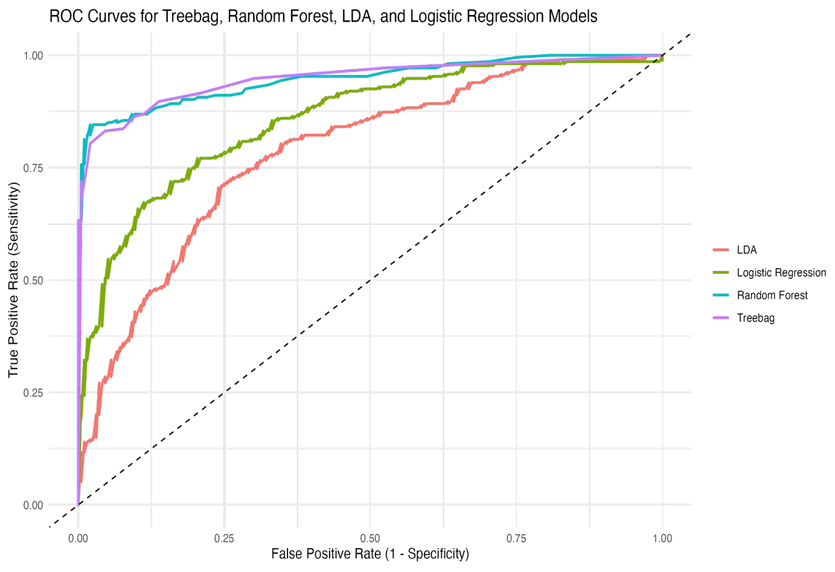

Figure 3:ROC Curve of Four Tuned Models, Severity Levels from 1 to 4

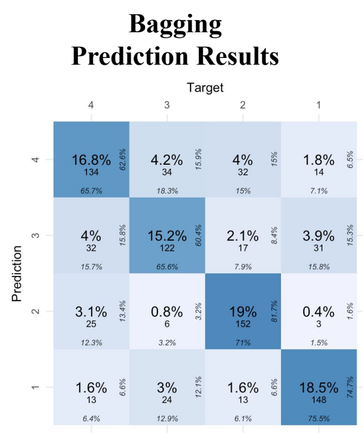

Figure 4: Confusion Matrix of Bagging Model

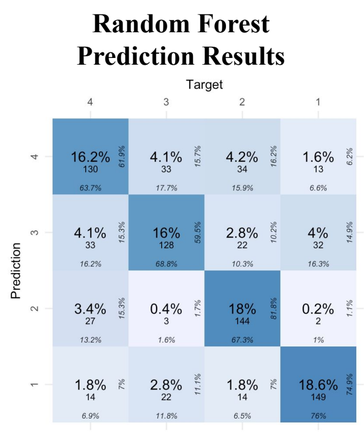

Figure 5: Confusion Matrix of Random Forest Model

Step 3: Train and tuned top 2 classifiers (bagging and random forest) on the entire training data set and we used the models used for prediction of entire testing set to get the final prediction results.

Points about Features Processing:

Feature	Process
State	- Separated State variable to better connect state location to level of severity. - Calculated probability of each severity level given state id.
Road signs near the accident	- Performed feature selection using the Boruta algorithm to focus on road signs accounting for most variability in severity.
Weather Condition	- Resigned weather condition into less groups to reduce tiny categories. - Performed feature selection using the Boruta algorithm to focus on weather condition accounting for most variability in severity.

RESULTS

Final Results & Comparison

- Used bagging , random forest ,LDA,Bagging models (tuned and trained on entire training data set) to predict accident severity levels of source testing data set.

Tuned model	Parameters	AUC
Logistic Regression	• Alpha=1 • Lambda=0	0.8572
LDA	• shrinkage=0.4	0.7847
Random Forest	• mtry= 33. • ntree = 200 • nodesize = 1	0.9486
Bagging	• minbucket = 2 • maxdepth = 5 • minsplit = 5	0.9482

Figure 6: Result of Tuned Model

The Figure 6 shows the categories of the adjusted four models, the specific adjusted parameters and the optimal values. According to the above parameters, the trained model is predicted on the test set, and the prediction result is obtained. In the indicator column, select the AUC value of the model in the test set and display it as a percentage.

In order to comprehensively compare the fitting and classification effects of each model, we chose to use AUC (Area Under The Curve) for the final model comparison, as can be seen from the above figure. After adjusting the parameters and compare the the model measurements such as AUC, accuracy and confusion matrix, **tree bagging became the slightly better model** compare to random forest and other models.

DISCUSSION

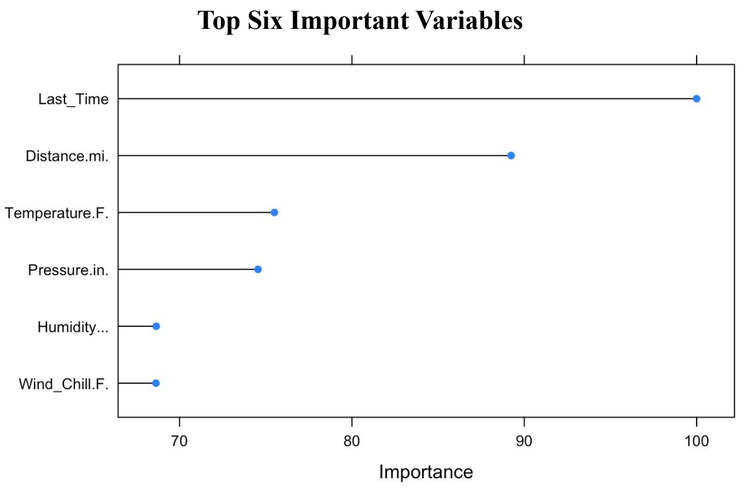

Figure 7: Top Six Significant Variables in Bagging Model

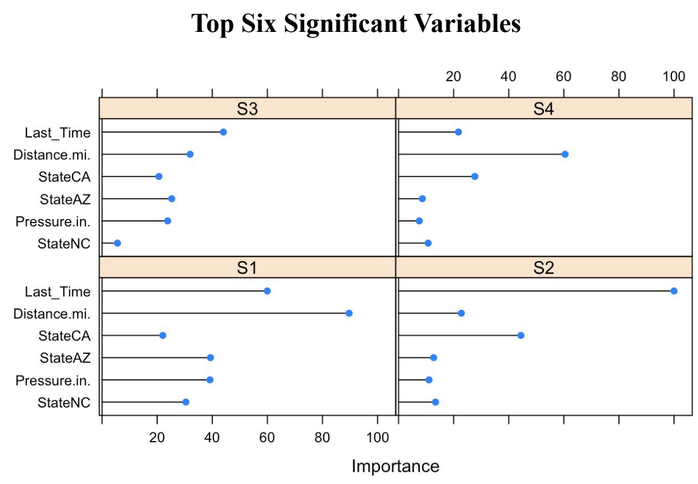

Figure 8: Top Six Significant Variables in Random Forest Model

In our analysis using bagging and random forest models, we identified duration, duration distance, and air pressure as highly important variables in both models. Notably, the random forest model also highlighted state id, CA, AZ, and NC as the top six important variables, given their distinctive characteristics in relation to the severity of car accidents. However, we observed that duration was less critical in predicting the extremely severe car accidents. This suggests that, in such cases, traffic jam time is not a significant predictor, possibly because these accidents involve a larger number of police rescue personnel, and rescue times are relatively faster.

In the bagging model, we found temperature, humidity and wind chill to be a significant predictor of the severity of car accidents. They suggest the weather factors, play an important role in predicting the likelihood and severity of car accidents, which can affect people from psychological and road conditions reasons. For example, humid weather and high air pressure can make people restless, while cold weather may indicate icy and slippery roads. The specific impact of these factors requires us to further explore the model.

CONCLUSION

In conclusion, our multi-class classification study has demonstrated that bagging and random forest models perform exceptionally well in predicting the severity of accidents in the U.S.Through the analysis of variable importance, we have identified key factors such as Last Time, Distance, Temperature, Air Pressure, Wind Chill, and Humidity as crucial determinants of accident severity. On the other hand, the presence of rare road signs showed minimal predictability in the context of accident severity.

These findings hold significant implications for policymakers and law enforcement officials, enabling them to focus their efforts on improving road safety, reducing accident frequency and severity, and enhancing rescue and traffic clearance operations. By understanding the impact of weather conditions on accident severity, traffic personnel can develop effective warning systems and prepare first aid measures in high-risk situations. Additionally, this study highlights the need for state transportation departments to prioritize traffic regulations, law enforcement, and increased patrolling in areas prone to severe accidents.

Future research should aim to expand the dataset we utilized, incorporating additional years and considering factors such as the COVID-19 pandemic, in order to provide an even more comprehensive understanding of the factors influencing accident severity in the United States.

References:

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New