Road Accident Severity Classification using US Accidents Dataset

Abdirashid Dahir, Department of Geography, The Ohio State University

***Abstract*—Most employees started to work from home due to social-distancing measures imposed by public health authorities to help prevent workplace exposure at the beginning of COVID-19 pandemic. As a result, gridlocked roads emptied out, and the congestion declined very sharply [1] . In order to predict accident-induced congestion severity levels, I utilized a huge US accident dataset of 1.5 million observations. Next, I predicted the accident severity classes using Random Forest (RF) Bootstrap Aggregation, and heuristic Support Vector Machine (SVM) - one-vs-one and one-vs-rest - after feature selection analysis (correlation coefficients and mutual information criteria). I then assessed the performance of classifiers through credible interval determination and binomial significance tests. The RF (bootstrap aggregation) outperforms both the base model (logistic regression) [2] , and heuristic SVM in terms of overall prediction accuracy, and confusion matrix metric. The study also demonstrates that traffic accident-induced congestion has been less severe than pre-pandemic levels.**

I. Data

The data ‘US-Accidents’ is a countrywide traffic accidents dataset that covers 49 states of the US [3][4]. The dataset was collected over a period of 4 years (2016-2020) using multiple APIs that stream real time traffic events captured by a variety of entities such as the US Department of Transportation, and law enforcement agencies through traffic cameras and traffic sensors.

The original dataset has 47 attributes that can be categorized under traffic (severity, accident start time, accident end time & distance affected); geography (street, city, county, zip code, state); weather (temperature, wind, humidity, pressure, and precipitation); Point of Interest (POI) such as cafes and train stations; and time of the day (sunrise, sunset, civil twilight, nautical twilight, and astronomical twilight).

In this study, accident-induced congestion severity classes (low:2, medium:3 & high:4) were the target variable while the rest of variables were used as feature variables. After encoding all categorical variables into one-hot numeric array, I generated 207 variables for subsequent feature selection.

II. Feature Selection

Script: `main.py`

I implemented a feature selection method to reduce the number of features in ’US-Accidents’

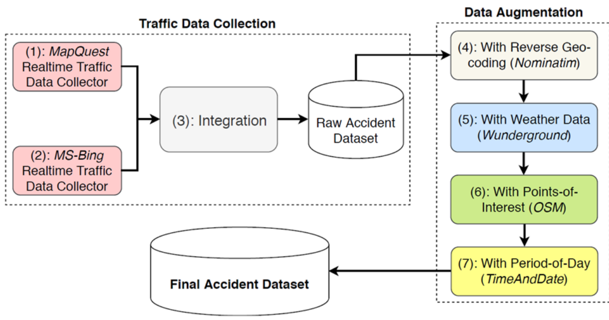


Fig. 1. The Process of constructing US-Accidents Dataset[3]

dataset before predicting accident severity classes. In order to reduce the computational cost of predictions, and improve the performance of the classifiers, I used a statistical-based feature selection method that involves understanding the relationship between each feature and the target variable. I selected the features that had the strongest relationship with the target variable based on Pearson's Correlation matrix and mutual information (information gain) from the field of information theory.

A. *Correlation Matrix*

I used Pearson's Correlation method for the 900,000 observations retained after the data cleanup procedure explained in [2] to understand the relationship among numerical features. I selected the predictor features using the following Pearson's correlation threshold:

$$X = \{X_k \text{ such that } |corr(X_k, y))| > 0.05$$
$$\text{where } k \in \{1, 2, 3, \cdots, K\}\}$$

B. *Mutual Information*

While Pearson's Correlation coefficient can quantify linear relationships, it fails to describe the dependence among variables that are related in a nonlinear sense. Therefore, I used information theory concepts like mutual information to explain the dependence among variables.

Let (X, Y) be a pair of random variables with values over the $\mathcal{X} \times \mathcal{Y}$. Let $P_{X,Y}$ be the joint distribution and P_X and P_Y be the marginal distributions. Mutual information is a measure of dependence that quantifies the statistical distance between the joint distribution of supposedly dependent variables and the product of their marginals, hence quantifying the mutual dependence between two variables. The formula for mutual information is given below:

$$I(X;Y) = \mathcal{D}_{KL}\left(P_{X,Y} \| P_X \otimes P_Y\right)$$

Finally, I selected 29 features that share high mutual dependence with the target variable for predictive modeling.

III. Models

Script: `main.py`

A. *Random Forest [Bootstrap Aggregation]*

Random Forest (bootstrap aggregation or bagging) is ensemble machine learning algorithm that is superior to bagged decision trees. I carried out sampling with replacement to reduce the variance of the forest (multiple decision trees) without increasing the bias. The RF algorithm showed better performance with 0.647 than logistic regression with 0.589, SVM (one-vs-one) with 0.54, and SVM (one-vs-rest) with 0.54. It also classified accidents more accurately than logistic regression classifier as shown in the figures below.

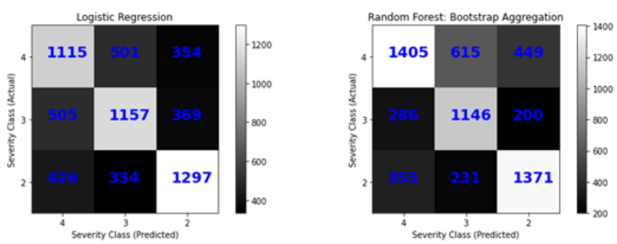


Fig. 2. Random forest versus base model (logistic regression)

B. *Support Vector Machine [one-vs-one and one-vs-rest]*

Support Vector Machines (SVM) are designed for binary classification problems. Therefore to overcome the inherently binary nature of SVM algorithm, heuristic methods (one-vs-one and one-vs-rest) are used to split up multi-class [accident severity classes: low(2), medium (3), high (4)] into different binary classification problem. Unlike one-vs-rest, one-vs-one splits the dataset into one dataset for each class versus every other class. Both heuristic SVM (one-vs-one and one-vs-rest) methods have achieved an overall accuracy of 0.554, and performed poorly in comparison to RF and base model (logistic regression) [2].

IV. Statistical Tests and Algorithm Comparison

Script: `main.py`

A. *Credible Interval Determination*

I computed the credible intervals of test accuracy sampled from beta distribution (1000 samples) for all classifiers.The 95% credible interval for test accuracy is 60% to 1.2% for all classifiers. This shows that the central portion of the posterior distribution contains 95% of scores between these two values. I also determined credible intervals for the overall accuracy of producer's accuracy of each severity class classified by all algorithms: logistic regression (base model) with 60% to 1.17%, RF (bootstrap aggregation) with 65% to 1.18%, RF (AdaBoost) with 66% to 1.2%, and neural network (3 hidden layers and 50 hidden units) with 63% to 1.24%. It is demonstrated that RF (AdaBoost) has the best overall accuracy.

B. *Binomial Significance Test*

Another statistical test based on target predictions for independent test sets is binomial significance test. Classifiers are compared to check if a new classifier is better than the old one. I ran accident severity classifiers on a test set to compare their accuracy scores. The following results show the probability that the classifier B is better than classifier A.

TABLE I

PROBABILITY THAT CLASSIFIER B (ROWS) IS BETTER THAN CLASSIFIER A (COLUMNS). NEURAL NETWORK (3 HIDDEN LAYERS & 50 HIDDEN UNITS)

	Logistic	**Random Forest [BA]**	**Random Forest [AdaBoost]**	**Neural Network**
Logistic	Nan	0	0	0
Random Forest [BA]	1.0	Nan	0.116	0.975
Random Forest [AdaBoost]	1.0	0.997	Nan	1.0
Neural Network	1.0	0.033	0.0	NaN

V. DISCUSSION AND CONCLUSION

Script: `main.py`

Two machine learning algorithms: I developed Random Forest (Bootstrap Aggregation) and heuristic Support Vector Machines (one-vs-one & one-vs-rest) based on feature selection analysis to predict accident-induced congestion severity classes in 49 U.S. states. Since the target variable (severity) has imbalanced classes, confusion matrix for each classifier is considered a more reliable evaluation metric. In addition, a boolean 'pre-pandemic' that I used to identify traffic accidents before and after February 2020 demonstrated that accident-induced congestion has been less severe than pre-pandemic levels. The study contributes to both methodological frameworks for predicting imbalanced classes, and the literature on the effect of COVID-19 pandemic on urban transportation networks.

REFERENCES

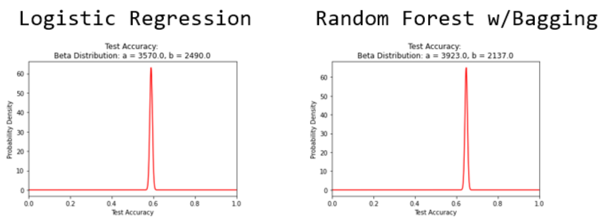
[1] J. Stiles, A. Kar, J. Lee, and H. J. Miller, "Lower volumes, higher speeds: Changes to crash type, timing, and severity on urban roads from covid-19 stay-at-home policies," *Transportation Research Record*, p. 03611981211044454, 2021. [Online]. Available: https://journals.sagepub.com/eprint/ETGWEHIYMYIAGJM5H6MT/full

[2] Y. Singh, "CSE 5523 project: Road accident severity classification using US accidents dataset," 2021.

[3] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath, "A countrywide traffic accident dataset," *CoRR*, vol. abs/1906.05409, 2019. [Online]. Available: http://arxiv.org/abs/1906.05409

[4] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident risk prediction based on heterogeneous sparse data: New dataset and insights," *CoRR*, vol. abs/1909.09638, 2019. [Online]. Available: http://arxiv.org/abs/1909.09638

APPENDIX




```
Overall Accuracy: 0.589
User's Accuracy: [0.566 0.57  0.631]
Producer's Accuracy: [0.545 0.581 0.642]
Kappa Coefficient: 0.383748
```

```
Overall Accuracy: 0.647
User's Accuracy: [0.569 0.702 0.701]
Producer's Accuracy: [0.687 0.575 0.679]
Kappa Coefficient: 0.470625
```

Fig. 3. Random forest versus base model (logistic regression)

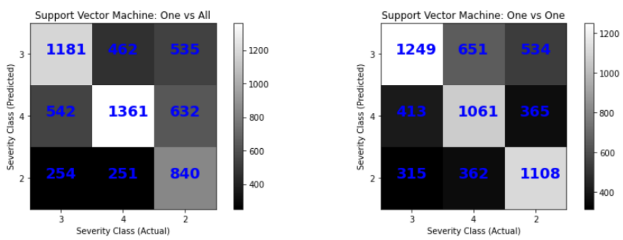


Fig. 4. Support Vector Machines (one-vs-one one-vs-rest)

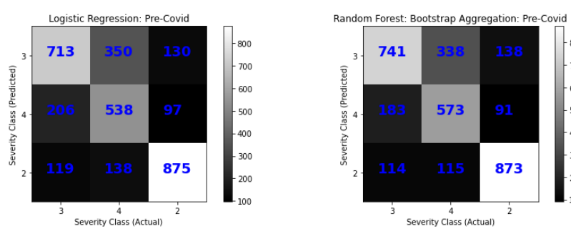


Fig. 5. Pre-Covid base model (logistic regression) versus Random Forest (bootstrap aggregation)

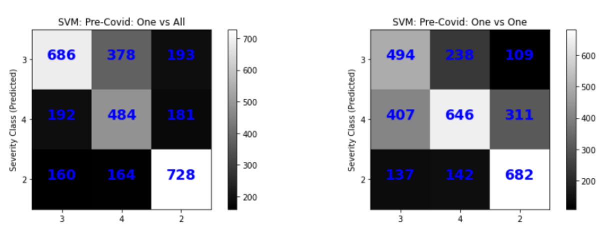


Fig. 6. Pre-Covid Support Vector Machine (one-vs-one one-vs-rest)

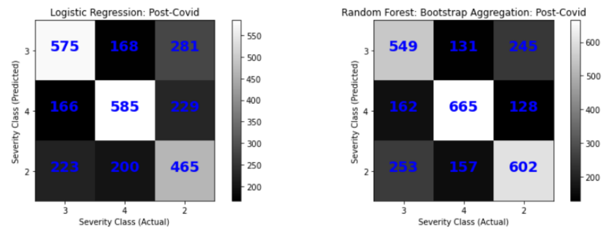


Fig. 7. During-Covid base model (Logistic Regression) versus Random Forest (Bootstrap Aggregation)

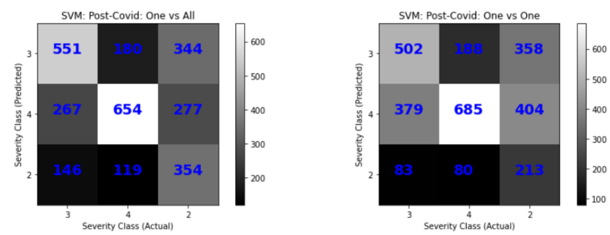


Fig. 8. During-Covid Support Vector Machine (one-vs-one one-vs-rest)

```
Logistic Regression
Credible interval for producer's accuracy
  Class: 2      Credible Interval:     mean: 0.7541912916086334 +/- 0.0112271794720363
  Class: 3      Credible Interval:     mean: 0.7331852481553098 +/- 0.0110065216064095
  Class: 4      Credible Interval:     mean: 0.7139947601164852 +/- 0.01203793705958988
Credible interval for overall accuracy  mean: 0.6006344797015292 +/- 0.0117916563926492

Random Forest: Bootstrap Aggregation
Credible interval for producer's accuracy
  Class: 2      Credible Interval:     mean: 0.7966461171571148 +/- 0.00975014436518306
  Class: 3      Credible Interval:     mean: 0.7856858742698303 +/- 0.01083297258111637
  Class: 4      Credible Interval:     mean: 0.7246280802706782 +/- 0.01179725721221880
Credible interval for overall accuracy  mean: 0.6534976788407888 +/- 0.01182832870894268

Random Forest: AdaBoost
Credible interval for producer's accuracy
  Class: 2      Credible Interval:     mean: 0.8049196608964672 +/- 0.01006639816651533
  Class: 3      Credible Interval:     mean: 0.77512498387612 +/- 0.01048705575034925
  Class: 4      Credible Interval:     mean: 0.7469166617444037 +/- 0.01116852045158800
Credible interval for overall accuracy  mean: 0.6633249373398463 +/- 0.01202217579279718

Neural Network: 3 hidden layers, 50 hidden units
Credible interval for producer's accuracy
  Class: 2      Credible Interval:     mean: 0.7708266958667148 +/- 0.01057496749825426
  Class: 3      Credible Interval:     mean: 0.763378277027271 +/- 0.0101825297223580
  Class: 4      Credible Interval:     mean: 0.726789105474825 +/- 0.0110922991777588
Credible interval for overall accuracy  mean: 0.6306008845891891 +/- 0.01240534218175437
```

Fig. 9. Credible Interval Estimations

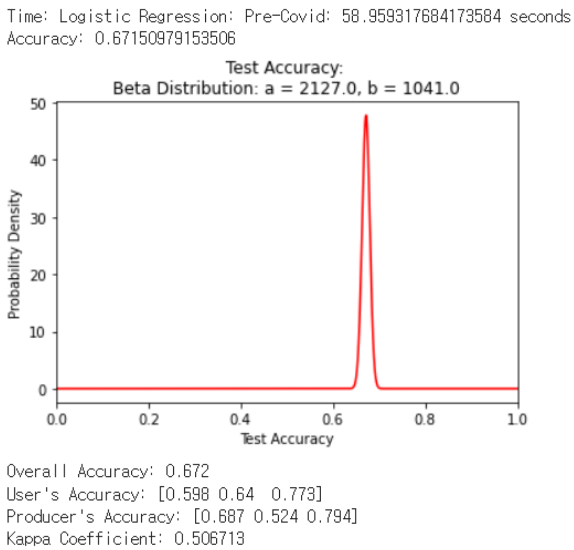


Fig. 10. Beta distribution for Pre-Covid base model (Logistic Regression)

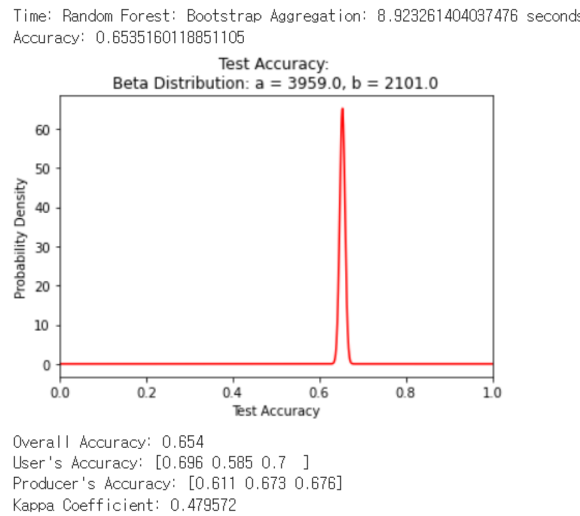


Fig. 12. Beta distribution for Random Forest (boostrap aggregation)

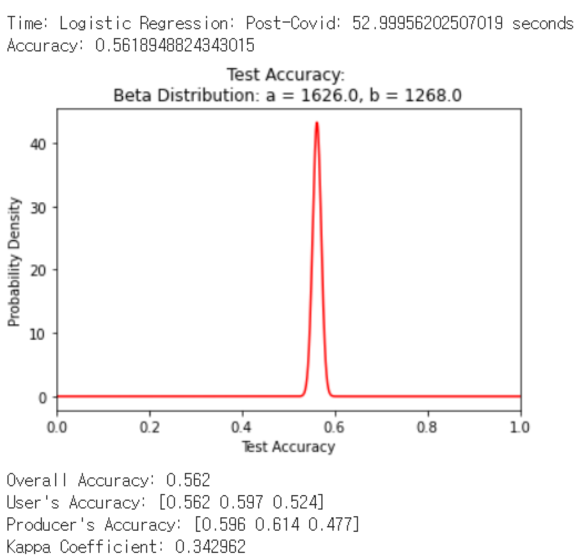


Fig. 11. Beta distribution for Pre-Covid base model (Logistic Regression

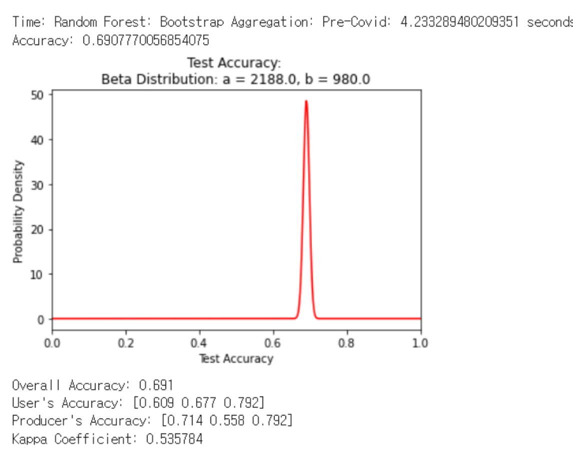


Fig. 13. Beta distribution for Pre-Covid Random Forest (bootstrap aggregation)

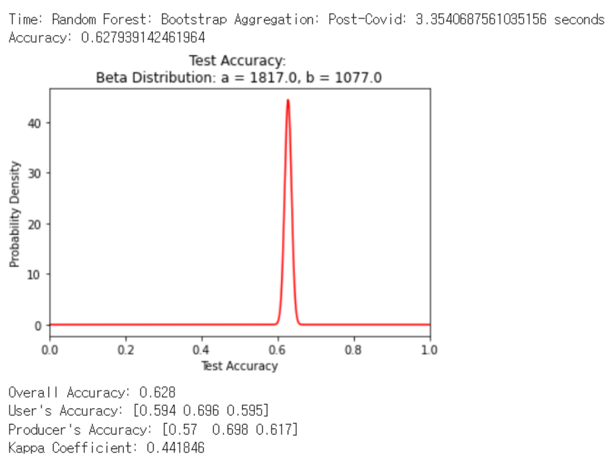


Fig. 14. Beta distribution for during-Covid Random Forest (bootstrap aggregation)