Predicting Fatal Aviation Accidents using Deep Learning

Qiaowei Jiang, Shiqi Ning, Guanzhi Wang, Xi Yang
Georgetown University Analytics Program

1. Abstract

This project researched on predicting fatal aviation accidents through various deep learning and neural network models. Multi-layer Perceptron Neural Network is used to predict the injury severity level (fatal or non-fatal). Dropout on the hidden layers of neural network is then researched and tested. The usage of dropout with proper dropout rate significantly ameliorate potential over-fitting problem. Super learner, which is a stacked ensemble model, is also applied to predict fatal aviation accident. However, super learner does not enhance prediction results and model robustness in this study.

2. Introduction and Problem Statement

Aviation transportation serves as a vital tool that provides people with mobility and access to various activities across the world.

Even though aviation transportation has been said to be the safest way of traveling, according to the US Department of Transportation (DOT) (2012), higher risk of safety can be posted due to various aspects:

1. **Human Behavior**: human error, fatigue, driver health, risky human behaviors such as drunk driving or being distracted by other means of communication
2. **Emerging and Automated Technologies**: fetal design of the public transportation tools that can cause fatality, and automatic pilot
3. **Weather-related Factors**: heavy rain and heavy fog.

Government agencies such as U.S. Department of Transportation (DOT) have already started safety data initiatives to identify safety challenges and find solutions that can save lives by integrating big data, and providing new insights.

This project will thus identify the factors that cause accidents, integrate them and find the weights of these factors by their importance in terms of severity level.

3. Related Work

Transportation safety has been of high priority for the society's public service system for a long time. The potential risks brought by safety uncertainty of the transportation system is of public's concerns. With recent progress of big data availability and machine learning technologies, more new methods of transportation safety analytics have been adopted.

Feature Selection and Sub-bagging

New method of accident severity analysis is applied in the imbalanced multi-class classification problem and, new quantitative method of predicting and assessing accident risks taking the relationship between travel time reliability and crash frequency are proposed by Miao (2018). Mabrouk (2016) applied artificial intelligence techniques as well as the development of several approaches and tools which assist in the modeling, storage and assessment of safety knowledge, to help with analysis and assessment of the safety of railway transport systems in France.

ShotSpotter (SST) and Domain Awareness System (DAS)

Machine learning is also the new trend for the government agency to do predictions. The NYPD has been applying machine learning methods to increase response to incidents since 2016, using the roll out of a system called ShotSpotter (SST), which processes sound from acoustic sensors placed in the streets to identify and locate gunfire incidents in real time (Lin and Rejniak, 2018). The integration of the SST sound database and the NYPD internal crime database, Domain Awareness System (DAS), promoting the way NYPD structured their operation and investigating gunshot (New York Police Department, 2016). Tools such as Hunchlab, or ShotSpotter Mission has been used to map each shift regions based on its risk assessment. And the predictive analysis (mass) implementation of identifying and mapping criminality profile of different regions has gradually become the new focus (Dubois, 2017).

4. Dataset

The public database for US transportation accidents or incidents is all provided by US government agencies including the Federal Aviation Administration (FAA), Federal Railway Administration (FRA), National Transportation Safety Board (NTSB). Through our research, NTSB provides the best database on transportation accidents or incidents.
Thus, data is downloaded from the NTSB aviation accident database, which records all accidents happened in the US, its territories and possessions, international waters, and other accidents investigated by NTSB (invited by other countries' government). The database can be queried by several filters and can be downloaded into XML format or delimited text format. The link to the database is attached below.
NTSB Aviation Accident Database:
https://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx

Data Preprocessing

- Only features of interest are preserved, including:
  - **1 response**: Injury Severity
  - **14 features**: Year, Month, Weekday, State, Investigation Type, Country Amateur Built, Aircraft Category, Aircraft Damage, Make, Broad Phase of Flight, Purpose of Flight, FAR Description, Weather Condition
- 'Unavailable' and empty data from all columns are removed.
- Year, Month, Weekday information are extracted from 'Event Date'.
- State information are extracted from 'Location'.
- Punctuations in the dataset are replaced with underscores.
- Empty 'Aircraft Category' are filled with 'Unknown'.
- Aggregated 'Make' into 21 groups.
- Only Title 48 of the Code of Federal Regulations in 'FAR Description' are preserved.

5. Methods

Multi-Layer Perceptron Neural Network with Dropout:

The first method we used is traditional multi-layer perceptron neural network with dropout layer after each dense hidden layer. According to the size of the dataset we have, we propose to use 2 hidden layers: first layer with 20-64 neurons, and second layer with 10-

32 neurons. Moreover, the dropout layer after each layer are added and dropout rate between 30% to 70% are tested. The activation functions in the hidden layer are RELU; the output function is sigmoid function; loss function is binary cross entropy; Adam algorithm is used as the optimization method.

Super Learner

Super learner model which is also known as stacking ensemble is used to enhance base machine learning model performance on regression or classification problem. Super learner algorithm is proposed by Polley and van der Laan (2010) as a powerful tool for machine learning and firstly applied in biology domain for gene expression research. Super learner uses any machine learning algorithms as base learners, and use an output function, which is usually linear, to generate final output result. The base learner chosen in this study are: MLP classifier, random forest, LightGBM, and XGBoost. The schematic of super learner is shown in Fig. 1 below.

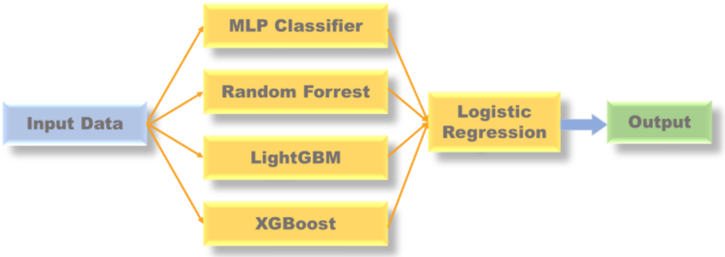


Figure 1 – Super leaner algorithm schematic.

6. Results

Figure 2 is the accuracy of loss function versus number of iterations applying Keras Sequential Model without dropout layer. Though the model performs well for train dataset, it exists significant overfitting problems, the loss function drops abruptly with the number of iterations increasing.

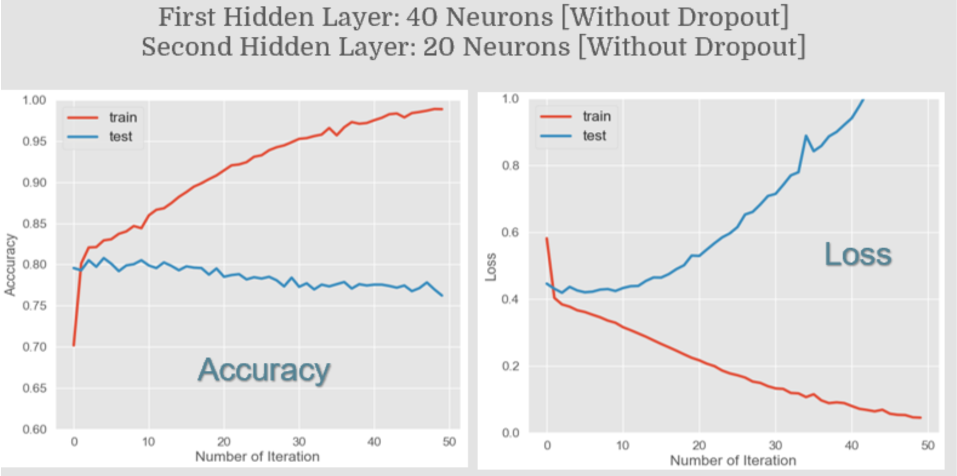


Figure 2 – Multi-layer perceptron neural network without dropout layer.

Figure 3 is the accuracy of loss function versus number of iterations applying dropout neural network based on the first model. It is obvious that dropout can solve overfitting problems efficiently. The performance for the train data and test data is approaching. Figure 4 shows the ROC Curve of train and test data, which also shows that there is no significant difference of ROC regarding train and test dataset and the accuracy for test dataset is above 0.8, which is satisfactory. In Figure 4, the test results, including confusion matrix and summary statistics, using the best neural network model with dropout layer is also shown. Regarding, precision, recall, and f1-score, our model all reached about 0.8 accuracy.

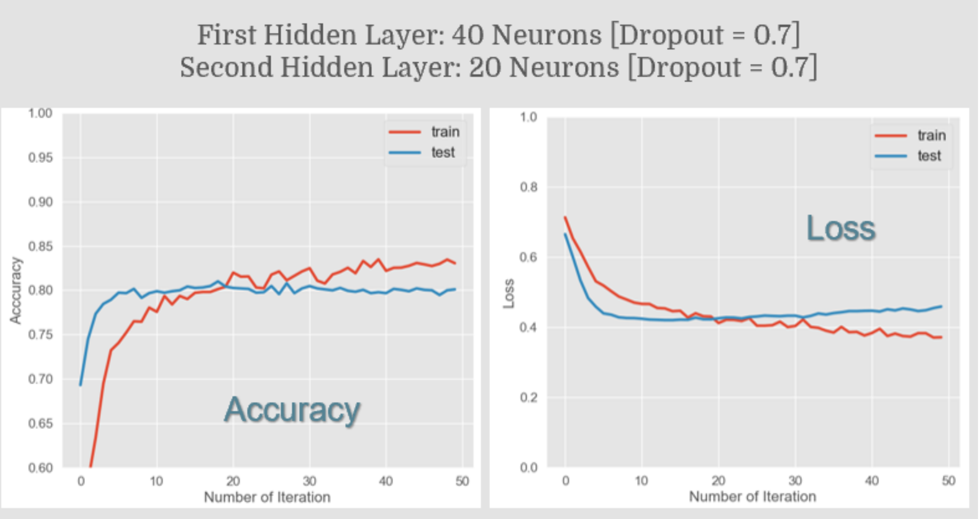
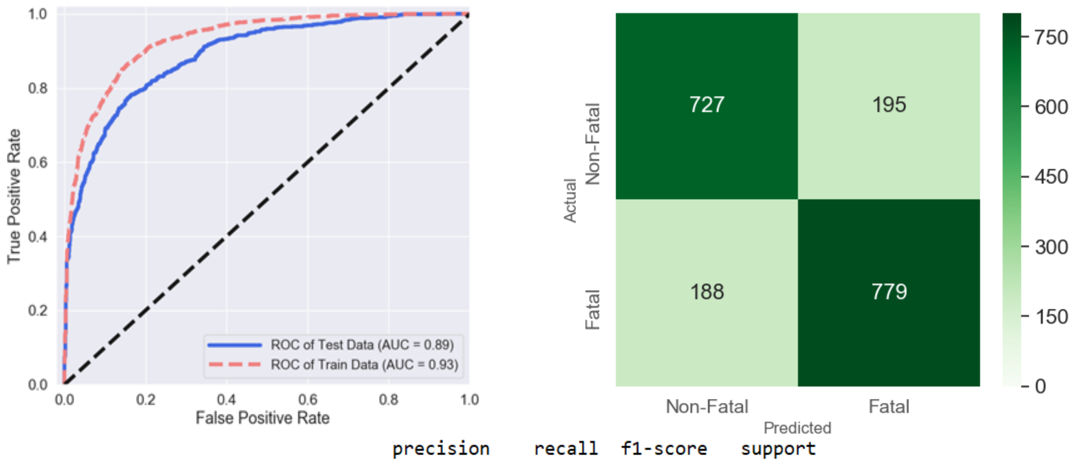
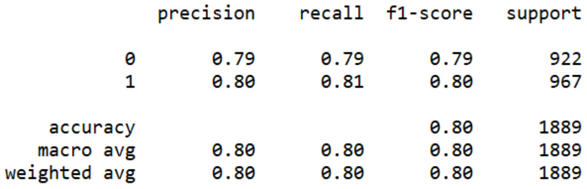


Figure 3 - Multi-layer perceptron neural network with dropout layer after each dense layer.





```
              precision    recall  f1-score   support

           0       0.79      0.79      0.79       922
           1       0.80      0.81      0.80       967

    accuracy                           0.80      1889
   macro avg       0.80      0.80      0.80      1889
weighted avg       0.80      0.80      0.80      1889
```

Figure 4 – ROC curve of train and test results, confusion matrix on test data of our best neural network model with drop out layers after each hidden layer, and test data summary report using the best model.

The super learner model does not yield "super" results in this study. In Fig. 5 below, ROC curves of super learner and its base learners on test dataset are shown. Super learner does not yield better results than the best base learner. The ROC curve of super learner is as good as the best base learner. Though super learner does not help in this study, in out opinion, we still expect to research more on this method in the future.

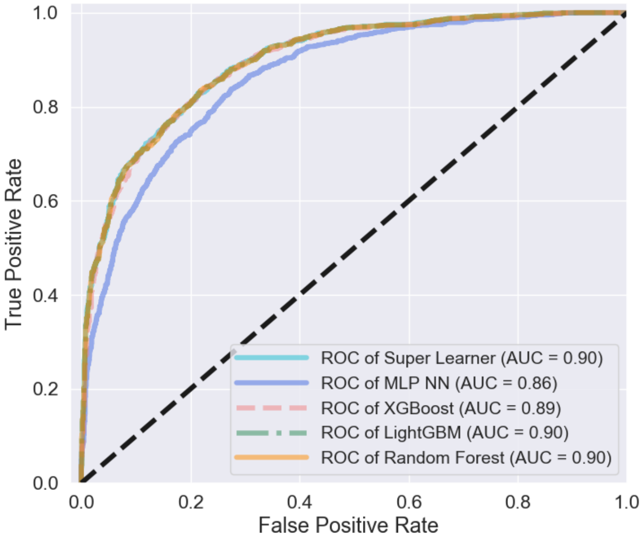


Figure 5 - Test data ROC curves of super learner and its base learners.

7. Discussion of Results

Since there are so many different influencing factors for transportation safety, the general model may exist overfitting, so it is innovative for us to apply dropout neural network to prevent overfitting. However, dropout model is difficult to interpret, so we refer to the related work, using feature selection model to help with explaining the important influencing factors.

The models applied in related work only take no more than 10 variables into consideration, and we include 14 influencing factors in our model, which is more consistent with the reality.

8. Conclusion

Dropout (neural networks) significantly helps improve overfitting problem in our case. From the visualizations shown in the above sections, the loss between the trani and test set significantly reduced when dropout is involved. Dropout guarantees the overall accuracy level of the model, it also allows training with the more important hidden nodes. However, compared to other advanced machine learning methodologies such as Super Learner, LightGBM, and Random Forest, Multilayer Perceptron Neural Network has a lower accuracy of 0.86, while the other methods reach an accuracy of 0.9. Even though, the difference is not significant, this may further indicate that the MLP NN is not suitable for data that contains multiple categorical features.

References

[1] Miao, Z. (2018). Transportation Safety Analytics with Statistical Machine Learning. Retrieved from https://repository.arizona.edu/handle/10150/631360

[2] Machine learning from experience feedback on accidents in transport - IEEE Conference Publication. (2016). Retrieved from https://ieeexplore.ieee.org/document/7939874

[3] Tami Lin, Malgorzata Rejniak. (2018). Smarter New York City: How City Agencies Innovate. Columbia University Press

[4] New York Police Department.(2016). The way moving forward. NYPD Publication

[5] Chantele Dubois.(2017). The future of AI and predictive Policing. Retrieved from https://www.allaboutcircuits.com/news/the-future-of-ai-and-predictive-policing/

[6] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[7] Polley, E.C. and van der Laan, M.J. (2010). Super Learner in Prediction. U.C. Berkeley Division of Biostatistics Working Paper Series: Working Paper 266.