Maximilian Steindl

Comparison of Algorithms for Predicting Traffic Accidents Severity in France

ABSTRACT

Road accidents are among the most critical challenges that is facing humanity as they lead to many deaths, injury, economic losses and fatalities each year. There is an overwhelming need for accurate models in transportation industry and government. This investigation was conducted to help all emergency responders in France to predict the severity of an accident based on the information given and to know what kind of resources (money, labor and equipment) needed to be deployed. There are many machine learning techniques that can be applied. This project used Logistic Regression (LR) and Random Forest (RF) on 2005 - 2016 France traffic accident data. The findings of this investigation show that RF can be very promising in predicting accident severity. RF has shown predicted performance with 71.6% than that of LR with 65.1%.

INTRODUCTION

Road accidents leads to deaths, injury, and property damage resulting in huge loss at both social and economic levels. According to research conducted by World Health Organizations (WHO) in 2018, approximately 1.35 million die each year as a result of road accidents. Moreover, road traffic accidents injuries are the leading cause of death for children and young adults between age 5 and 29 years. Furthermore, road accidents cost most countries about 3% of their gross domestic product (GDP). The 2030 Agenda set by WHO is geared toward cutting the number of accidents by half.

As the demand of vehicle increases the world today the number of injury and death from traffic accidents is projected to go up. According to HAL archive of 2014, road accidents costed France about 35.7 and 50 billion Euros, and that number is expected to rise. Clearly there is a problem in transport systems that needs to be solved, thanks to Machine Learning Algorithms and Big Data tools that can readily be applied to collect this information that be useful in helping solve these challenges.

Supervised learning techniques and to be specific classification methods are among commonly used methodology to mine traffic data. These methods are trained on known datasets where factor that contribute to accidents are known. These machines then are used to make smart decisions to predict accidents severity in real time and which in turn can eradicate avoidable accidents.

The goal of this project is to achieve accuracy in predicting traffic accident severity by studying known factors. These include road category, traffic regime, number of traffic lanes, surface

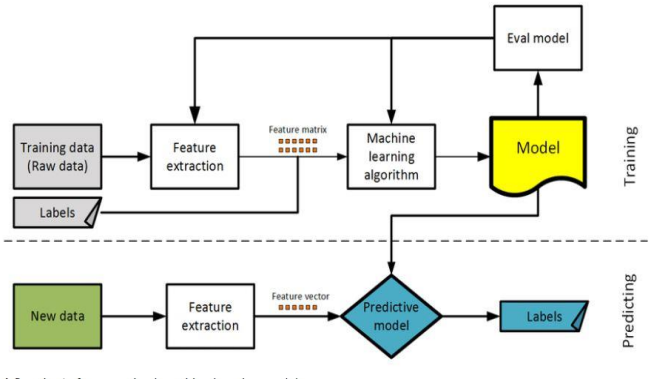
condition, and infrastructure to predict “grav” which stands for severity in the known dataset. Additionally, this study aims at helping department of emergency to allocate resources efficiently and thus save time, money and proper allocation of equipment.

METHODOLOGY

The purpose of this section is to go over the methods used during this research to build the prediction classification rules of the best performing model (LR and RF).

Figure 1. Model Development Diagram

a. Data source



The data used for the purpose of this study was acquired from Kaggle.com open source for data distribution center. The dataset contains information about road crashes for France from 2005 through 2016. The data contains 839,985 of recorded crashes.

Accident severity data was categorical as follows: 1: unscathed, 2: Killed 3: Serious injury and 4: Light injury as shown in Figure 2.

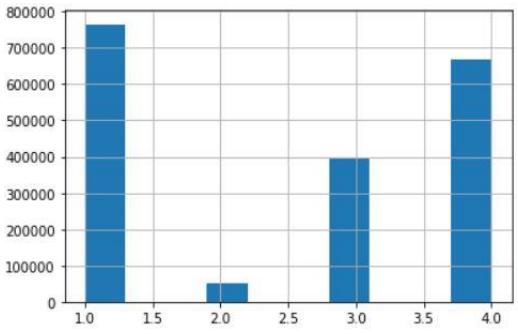


Figure 2: Distribution of Severity

Over 50,000 people lost their lives during the over 10 years period accounting to about 2.70 % of mortality from crashes in France.

b. Data Pre-processing

Preprocessing of data was performed before model training and testing. These processes included cleaning (data wrangling), normalization, feature selection(extraction), and transformation. The dataset used integer values for entire attributes. Transformation was done on categorical data to contain 0s and 1s. Missing values were accounted for by mean (average) and frequently occurring integers. Factors that affected severity of accidents were chosen based on importance as follows in Table 1.

feature	importance
dep	0.334
jour	0.271
hrmn	0.151
mois	0.096
catr	0.091
agg	0.057

Table 1. Feature Selection Table

c. Machine Learning Algorithms

To determine the effectiveness of machine learning technique needed for classification, the dataset was split into a training set and testing set. The purpose of doing this was to make sure that the models are not familiar of the dataset that are going to be tested on in order to avoid overfitting. For this study, the dataset was split into 80% training and 20% testing. The next step was building the model for predicting severity. There are a variety of ML Algorithms to choose from but for this study RF and LR were selected.

Random Forest (RF) is an ensemble learning methodology which constructs a series of decision trees during training and results into classes(classifications) or mean prediction of individual trees. Logistic Regression (RF) on the other hand is a classification algorithm that maps results of linear function to a sigmoid function. It very simple and yet powerful.

Other ML Algorithms that are of interest to researchers are Naïve Bayes Classification (NBC) and AdaBoost Classification.

d. Performance

Three performance measures were used to compare the two machine learning algorithms of interest. These were Accuracy, Jaccard score, f1-score and log loss.

Model	Logloss	Jaccard	F1Score	ROC
LR	0.642	0.367	0.639	0.651
RF	0.914	0.440	0.665	0.71.2

Table 2: Results of Performance Measurement

RESULTS AND DISCUSSION

This section will go over the results of the two classifiers RF and LR.

a. **Accuracy**

Recall is a measure of quantity or completeness or quantity while precision is a measure of exactness or quality. High recall means that the model returned relevant results and high precision means that the model returned more relevant results than irrelevant. Log loss is a measure of how likely an event will occur. Jaccard is a measure of how similar or dissimilar the dataset is to each other.

The bar chart in Figure 3 compares RF and LR ML algorithms, from the results of F1-score, RF performed very well with a precision of 66.5% compared to LR with 63.9%. It is also clear from the results that RF can performed better in showing how similar the dataset was with 44.0% accuracy compared to LR with 36.7%. Furthermore, RF had 91.5% accuracy of predicting severity of an accident.

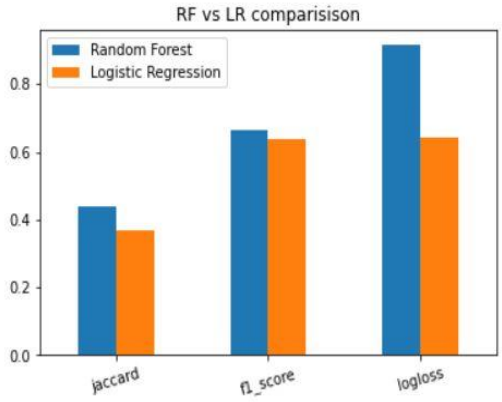


Figure 3: Jaccard, F1-score, Log Loss

a. **The Area Under ROC Curve (AUC)**

The value under the curve is between 0 and 1. A value close to 0.5 and below concludes a bad model. Excellent results start at 0.7 and above. Now, lets analyze the results of Figure 4 below.

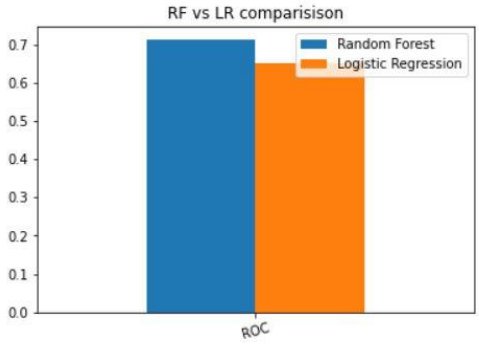


Figure 4. Area Under ROC Curve

Again, from the observation one can conclude that RF performed way better with 71.6% accuracy than LR with 65.1%. It is evident from the result that RF outperformed LR in all aspect and therefore the best classifier of the two.

SUMMARY AND CONCLUSION

This study was to investigate two ML algorithms that are prominent in solving classification problems, LR and RF. From all the results (jaccard, f1-score, log loss and ROC curve) one can conclude that LR is an excellent predictor of accident severity with an ROC value of 71.6%. Surely these results provide enough evidence that the transportation system and government of France and around the world can benefit from RF Classifier. When deployed this model can help on how to allocate resources (money, equipment and labor) in real time whenever an accident occurs. Now that there is a way to predict severity of an accident the emergency departments are now able to know whether to send a helicopter or just an ambulance, whether to send more help or just a handful and much more. This will help save lives, money and time.

REFERENCE

[1] "Road Traffic Injuries." *World Health Organization*, World Health Organization, www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[2] Mimi, A. (2018, June 13). Accidents in France from 2005 to 2016. Retrieved October 13, 2020, from https://www.kaggle.com/ahmedlahlou/accidents-in-france-from-2005-to-2016

[3] Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Incorporated.

[4] Carnis, L. (2018, May 16). *The cost of road injuries in France: some preliminary outcomes*. Archive Ouverte HAL. https://hal.archives-ouvertes.fr/hal-02265815/

[5] "Find and Share Research." *ResearchGate*, www.researchgate.net/.