## Introduction

*The goal of this project is to build a crash prediction model that identifies and explains the contributing factors of risk, and accident severity in order to be used to assess the impact of intervention for car accidents.*

Car accidents are still one of the biggest causes of death in the world. For that reason, the development of models for an efficient and accurate prediction of traffic accident location and severity, is necessary. This project analyzes circumstances of road accidents in the UK in 2017. The statistics relate only to personal injury accidents on public roads that are reported to the police, and subsequently recorded, using the STATS19 accident reporting form. This project is intended for City transportation departments and those responsible for managing risk on road networks.

I am replicating and adapting a proposed a methodology (Tambouratzis & Souliou & Chalikias, 2014) that involves feature selection of the accident parameters that maximize prediction accuracy (implemented through Random Forest search), followed by feature extraction through principal component analysis and selection of minimal components that contain relevant information of the original parameters. Finally, I applied accident severity prediction through random forests, and I compared with existing accident classification/prediction approaches in the field.

## Preliminary Results

The road accidents in the UK in 2017 contains 59554 data points, with each one consisting in 53 input parameters that have been collected through the police, and subsequently recorded, using the STATS19 accident reporting form. Through data pre processing, I dropped the rows that contain Null values, given that the data set is big enough to do not affect the results, and columns that presented similar information. It is necessary to examine whether dropping out the null values reduce the amount of number of classes accidents that could affect the results given their small presence. In this case, the percentage of fatal accidents were not affected, then I proceed to delete them. Given that the data is heavy imbalanced, with fatal accidents - which it is of paramount importance to predict- having a really small proportion in comparison with the other 2 classes. I balanced out this; however, for more relevant results a better data set with a greater number of fatal accidents (sadly, but necessary), will be ideal.

First, the most relevant risky features that cause or contribute to accident severity are determined and retained through Random Forest (RF). In the only case where RF does not perform well when features are monotonic transformation of other features. However, I selected this methodology given that as demonstrated in the correlation graphs, the data is not strongly correlated, just for a few cases that were manually discarded. In addition, for this case is a great option given that it handles the missing values and maintain the accuracy of a large proportion of data. This step improves efficiency by eliminating redundant or irrelevant data.

Second, a transformation of selected parameters into orthogonal components through PCA is implemented, I normalized the data previously. PCA is used to reduce the dimensionality such that a linear model no longer badly overfits the data. Using just the 10 best components that explains most part of the data, boost the efficiency of accident prediction. Finally, Random Forest, Logit Regression and Decision Trees are used for evaluating prediction accuracy comparing with and without the pre process through PCA, to analyze if this improves the accuracy of the model.

**Results and Suggestions**

None of the models presented improvements in their accuracy after the application of parameter selection followed by the PCA implementation. I ran the PCA without component constraints at first. This showed that really uncorrelated data presenting 10 components, from an input of 10 features. Which makes sense, given that these features are the most relevant according to the RF. When trained, the PCA + models showed very little improvement over the raw classifiers. In fact, the accuracy in some of the model even decreased. We can think about reasons why PCA did not present any improvement. First, given that the PCA uses feature-level correlation, there could be the case in which our data are correlated non-linearly, making the PCA perform poorly. Second, PCA, as an unsupervised model, its direction are selected to maximize variance, in order to present more information about the model. However, if this is not considering the class information it is possible that the result will present the different points of each class very close together. Separability is completely ignored in this method. PCA could miss key separation dimensions, creating a data hard to classify later. Further research is needed for more relevant conclusions.

One improvement could be identifying the most important features and concatenate them with the rest of features reduced by a PCA implementation. In this way, we could hold information about the ones we did not choose to include, but offering an improvement in the efficiency. Another suggestion, could be replacing or adding an LDA implementation. LDA, a supervised dimensionality reduction, maximizes separability of the classes under some very strict assumptions on the underlying data. However, again, further research is needed for more relevant conclusions.

The accuracy of the different models presented a acceptable score. Decision tree and Random Forest performed much better in terms of predicting all the classes of accident severity than the Logistic regression. This last one, presented a better accuracy but a poor classification in the three classes. Through the accuracy board in the Random Forest, we can observe a high accuracy in the training data, however, the fatal accidentes (class=1) are almost/ not included, except for the Decision Tree model. Therefore, one option can be do hot-encoding to re-labeled the Accident Severity between 1 or 0 float, representing Fatal or Nonfatal classes. Then I applied Logistic regression and Random Forests in order to measure the classification accuracy. This along with a more balanced data set, could improve the accuracy of the classification.
Other Variables to Consider

- (Estimated) speed of vehicle(s) at time of incident

- Distracted driving factors such as mobile phone use
- Driver intoxication
- Mountains, fields, or other road areas
- Presence of bicyclists on the road
- Human factors such as population density and distractions such as billboards
- Graph derived features of the road network such as centrality and flow
- traffic volume data to understand which roads experience the highest traffic and how changing trends of usage might affect risk
- more detailed road features including speed limits, signals, bike lanes, crosswalks, parking etc.
- road infrastructure data

**Resources**

Caliendo, C. (2005) Principal Component Analysis Applied to Crash Data on Multilane Roads. Retrieved from https://www.researchgate.net/publication/242220762_Principal_Component_Analysis_Applied_t o_Crash_Data_on_Multilane_Roads

Malik, U. Implementing PCA in Python with Scikit-Learn (2018) Retrieved from https://stackabuse.com/implementing-pca-in-python-with-scikit-learn/

Mujalli, R. & De Oña, J. Injury severity models for motor vehicle accidents: A review (2014). Retrieved from https://www.researchgate.net/publication/270466780_Injury_severity_models_for_motor_vehicle _accidents_A_review

Singh, R. UK Traffic Accidents (2017) Retrieved from https://www.kaggle.com/ambaniverma/uk-traffic-accidents

Tambouratzis, T. Souliou, D. Chalikias, M. (2014) MAXIMISING ACCURACY AND EFFICIENCY OF TRAFFIC ACCIDENT PREDICTION COMBINING INFORMATION MINING WITH COMPUTATIONAL INTELLIGENCE APPROACHES AND DECISION TREES. Retrieved from https://content.sciendo.com/view/journals/jaiscr/4/1/article-p31.xml