Hack the Crash!

Adell V., Autrand P., Rodríguez J., Rodríguez A.

HackUPC 2019

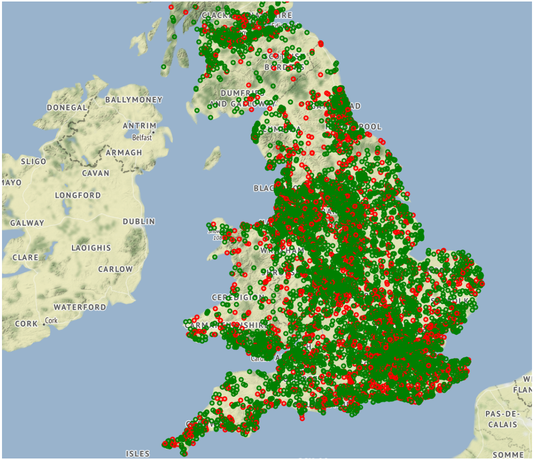
October 13, 2019

Overview

1. A brief on the approach
2. Data preprocessing and feature engineering
3. What does our final model look like?
4. Our takeaways
5. Actions to reduce the severity of accidents

A brief on the approach

1. Remove duplicates
2. Remove outliers
3. Encode cyclical features
4. Encoding categorical variables: Label and Target encoding
5. Data in variable amounts
6. Model search



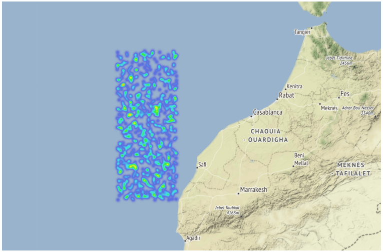
Data preprocessing and feature engineering

Remove duplicates

- 1024 rows in `accidents.csv`
- 4256 rows in `vehicles.csv`

Remove outliers

- 214 roads with a speed limit of 300mi/h
- 824 locations which are clearly not a UK overseas territory



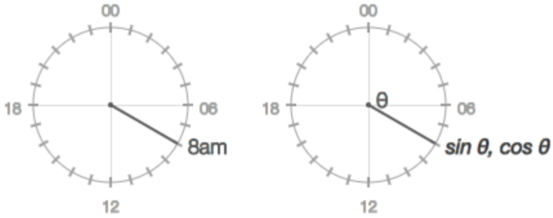
Note: We have found that there are 22512 **missing vehicle references exclusively on the test dataset.**

Data preprocessing and feature engineering (cont.)

Encoding cyclical features

Question: How do we tell the model that 24 = 0?

- We need a representation such that 24 and 0 and close together, and 24 and 12 are far apart
- Solution: Angles



Data processing and feature engineering (cont.)

Encoding categorical variables: Label encoding

Replaces strings with an index. Good for $=$ type questions.

However, we would like another representation that allows us to make $<$ or $>$ type questions.

Uncomparable variables

Consider the case of the variable `Junction control`. Examples of values it can take are `Auto traffic signal` and `Give way or uncontrolled`.

How can we compare them? A simple yet powerful approach is target encoding

Encoding categorical variables: Target encoding

Replaces a categorical variable with the conditional probability of the target given the variable

$$\texttt{JC} = \texttt{ATS} \implies P(\texttt{ACC} = \texttt{SEV} | \texttt{JC} = \texttt{ATS})$$

Data processing and feature engineering (cont.)

When we use vehicle data to predict accident severity, we run into an issue: the amount of vehicles differs from accident to accident.

Thought process

- We need an encoding that works with 1, 2 or any number of cars.
- With continuous variables, a good option is to use the mean, minimum or maximum. However our variables are categorical!
- It is not clear how to do this. However, we already know how to convert categorical into numeric variables: **using target encoding**.

Dealing with variable amount data

1. Use target encoding to convert categorical variables into numeric probabilities.
2. Extract the `mean`, `minimum` and `maximum` for each of these variables.
3. We end up with 3 new useful numeric variables that are representative of the set of cars, no matter the size.

What does our final model look like?

Classification trees are some of the best performing models out there.

Thought process

- Our set was predominantly categorical ($\sim 80\%$)
- It was much simpler to use trees as opposed to neural nets

Model search

After feature selection and extraction, we tested four major types of classification trees, along with several combinations of base classifier amount and probability cutoff.

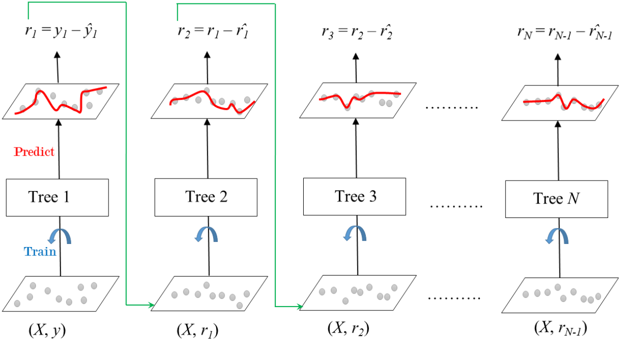
1. Vanilla Random Forest
2. Bagging
3. AdaBoost
4. Gradient Boosting

Gradient boosting came out on top: 0.42 F1 (on a validation split)

Gradient boosting

Gradient boosting algorithm

1. Ensemble method
2. Creates a sequence of weak learners that attempt to build on the weaknesses of the predecessor
3. The weak or base learners are classification trees



Our takeaways: What gets measured gets done

1. It is important to look at the data

Visually detect outlier data

2. Variables that really matter

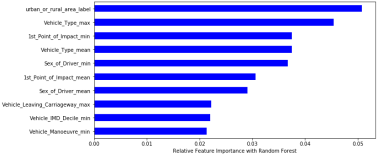
The vanilla RandomForest classifier, while not as good during prediction, has a feature importance function. Using this, we can get a pretty good idea of what has the biggest impact on accident severity.

1. Urban or rural area: Rural roads are more sever accident prone than urban roads. This is most likely due to their inferior conditions.
2. Vehicle Type "max": The worst vehicle type has the second highest influence on accident severity.
3. 1st point of impact "min": The safest point of impact has the third highest influence on accident severity.

3. We can measure categorical variables using target encoding

Combine and compare

Actions to reduce the severity of accidents



Possible actions

Based on the feature importance information, we can do the following to solve the problem:

- Urban or rural area: A detailed study is probably necessary, but a priori this would suggest that improving rural road conditions would have a big impact.
- Vehicle type "max": Not very much one can do about this beyond alerting drivers or sectioning off lanes for big vehicles, but perhaps this can have a sizeable impact.
- 1st point of impact: Advising vehicle makers on areas to reinforce, or research ways of reinforcing.

Possible business plan

General steps

Sell the tool as a resource management aide (SaaS).

- Governments and local authorities of the UK can subscribe to use our predictive model and feature importance insight to combat accidents (for instance, in the way we described in the previous slide).
- The model can keep improving if we continue to train on newly generated data. Even more so, the more clients we have, the greater the database and the better our product.
- Thus, we can continue the service as a continuous product rather than a one time sell (subscription based revenue stream).

Our product is attractive because car accidents are one of the major mortality causes in the modern day. Furthermore we offer clear solutions on how to address the problem, and offer insight on possible root causes.

Usage of software and IT infrastructure in our solution?

Limitations of our approach

- Our model was trained on a laptop
- Our computational and distributional resources are limited

With software and IT resources

- Train a bigger model (more weak learners), using greater memory and GPU access.
- Maintain a continuous service with auto-updating model on a dedicated server.
- We discarded the latitude and longitude because we decided that as numbers they wouldn't prove useful (beyond discarding car accidents in the atlantic).
- However, with a sufficiently good satellite, we could use the image of the accident site as valuable data.

Economic impact

Reduced and more efficient spending

It allows government and local authorities to use their resources more efficiently. This leads to reduced spending and greater value per euro.

Human impact

Moreover, there is also the human impact of our solution. By concentrating resources on policing and improving accident prone areas, we can prevent heavy injuries and deaths.

The combined effect of lives saved in the economy is surely important.

Other applications

As a consumer application

- We can use our model to warn consumers that driving through a particular route has a high probability of accident.
- Moreover, since our model makes use of driver data, we an personally tailor the "danger sensor" to every individual user.

Prevention is better than cure

- While what makes bad road is clear, our product can provide valuable information that helps city engineers know what **not** to build.

We can prevent "danger situations" by building "safe" roads.

Translating insights into technology improvements

Smart cities

- Smart cities can have sensors at accident prone junctions and roads
- Moreover, interest in developing smart cites in this area will naturally yield technology in others
- For example, accident prevention may stimulate automatic emergency services deployment and road sectioning.
- Improvement in those areas go beyond much more than our problem.

Thanks!