Accident analysis and severity prediction

The problem

- **Road accidents** are a huge problem and cost (human lives, money and resources) for modern society
- Estimating the likelihood of a road accident and its possible severity with a **predictive model** is thus a very important topic

This brief study will develop and test predictive models of accident severity, considering as input variables a set of conditions (road, time, weather, etc.)

Possible **stakeholders**:

- Governments, police forces, road authorities (determine effective countermeasures)
- Normal drivers, cyclists or pedestrians (warning about dangerous situations)
- Producers of driving assistance systems
- Self-driving systems (security protocols)
- Insurance companies (risk evaluation)

Data source

- Database of UK accidents collected by police forces https://data.gov.uk/
- Available under the Open Government Licence http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/
- For this study, only the 2017 and 2018 ‘accident’ databases are considered; the 2016 database will be used for model test

	Accident_Index	Location_Easting_OSGR	Location_Northing_OSGR	Longitude	Latitude	Police_Force	Accident_Severity	Number_of_Vehicles	Numb
0	2017010001708	532920.0	196330.0	-0.080107	51.650061	1	1	2	
1	2017010009342	526790.0	181970.0	-0.173845	51.522425	1	3	2	
2	2017010009344	535200.0	181260.0	-0.052969	51.514096	1	3	3	
3	2017010009348	534340.0	193560.0	-0.060658	51.624832	1	3	2	
4	2017010009350	533680.0	187820.0	-0.072372	51.573408	1	2	1	

5 rows × 32 columns

Strengths of the databases:

- data are rigorously validated;
- they are collected with the same methodology across the whole UK;
- they include a lot of details regarding the accidents.

Weaknesses:

they record only accidents involving:

- at least one vehicle
- in which at least one person was injured (no damage-only accidents)
- and which were reported to the police (possible undersampling of slight accidents)

Data selection and cleaning

- Original combined (2017+2018) database: 252617 rows, 32 columns; after data cleaning: 249441 rows (less than 1.5% rows lost)
- Target for the model: **Accident_Severity**: 1 *Fatal*, 2 *Serious*, 3 *Slight*
- 10 columns used for model training:

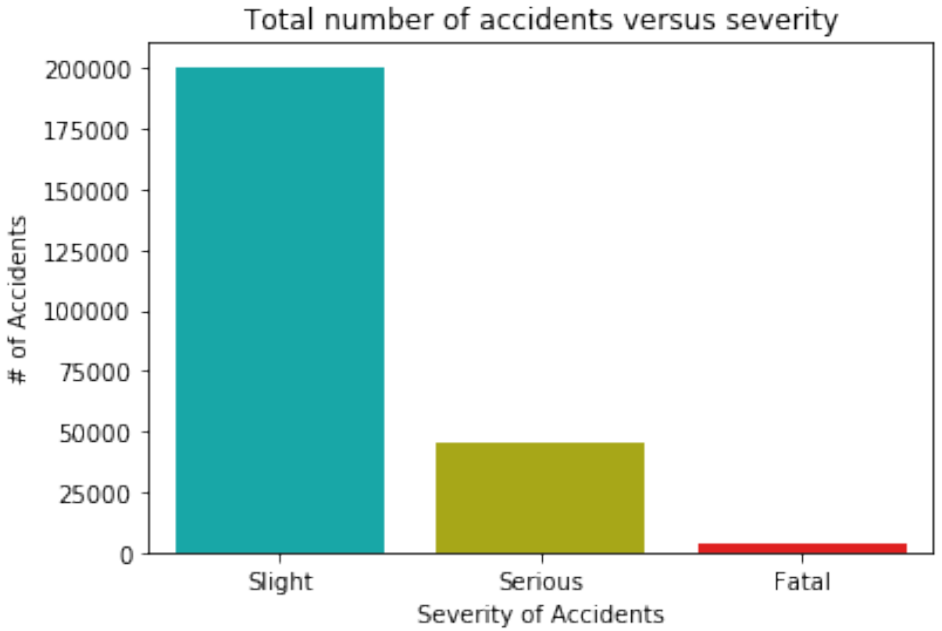
  - Day_of_Week, Time_Period, Season;

  - Road_Type, Speed_limit, Junction_Detail;

  - Light_Conditions, Weather_Conditions;

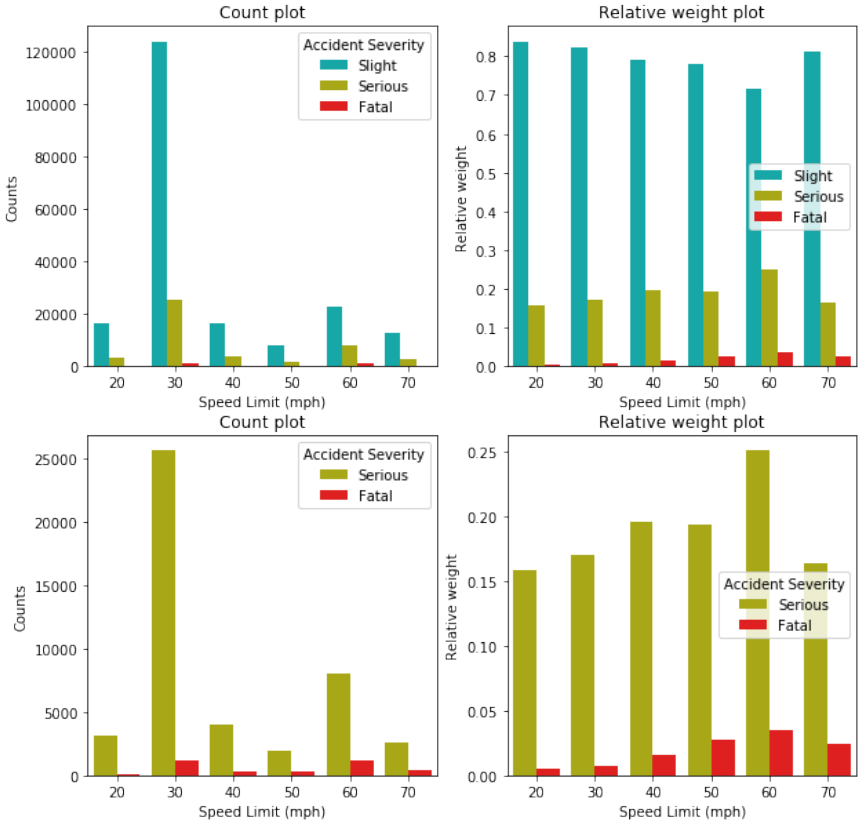
  - Road_Surface_Conditions, Special_Conditions_at_Site

Preliminary evaluations



- 80.4% slight accidents
- 18.2% serious accidents
- 1.3% fatal accidents

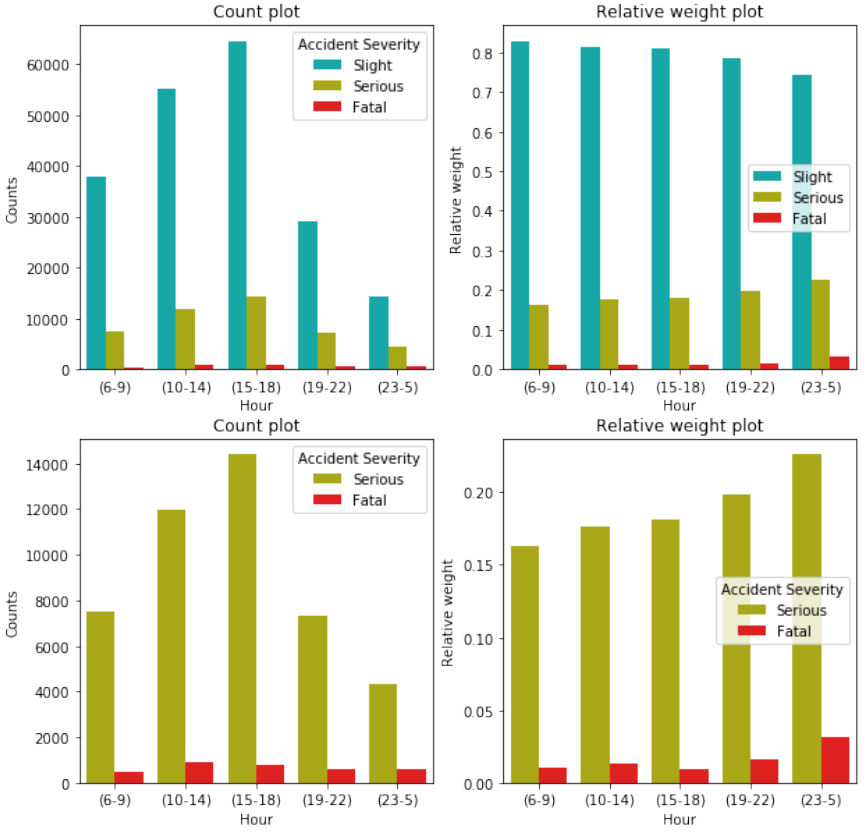
- The severity distribution is **highly unbalanced** towards slight injury accidents
- It is very important to take into account this when building the prediction model.



- Left: Count distribution
- Right: Relative Weights (to counts at that particular x-axis value)

In the bottom row, only serious and fatal cases are shown

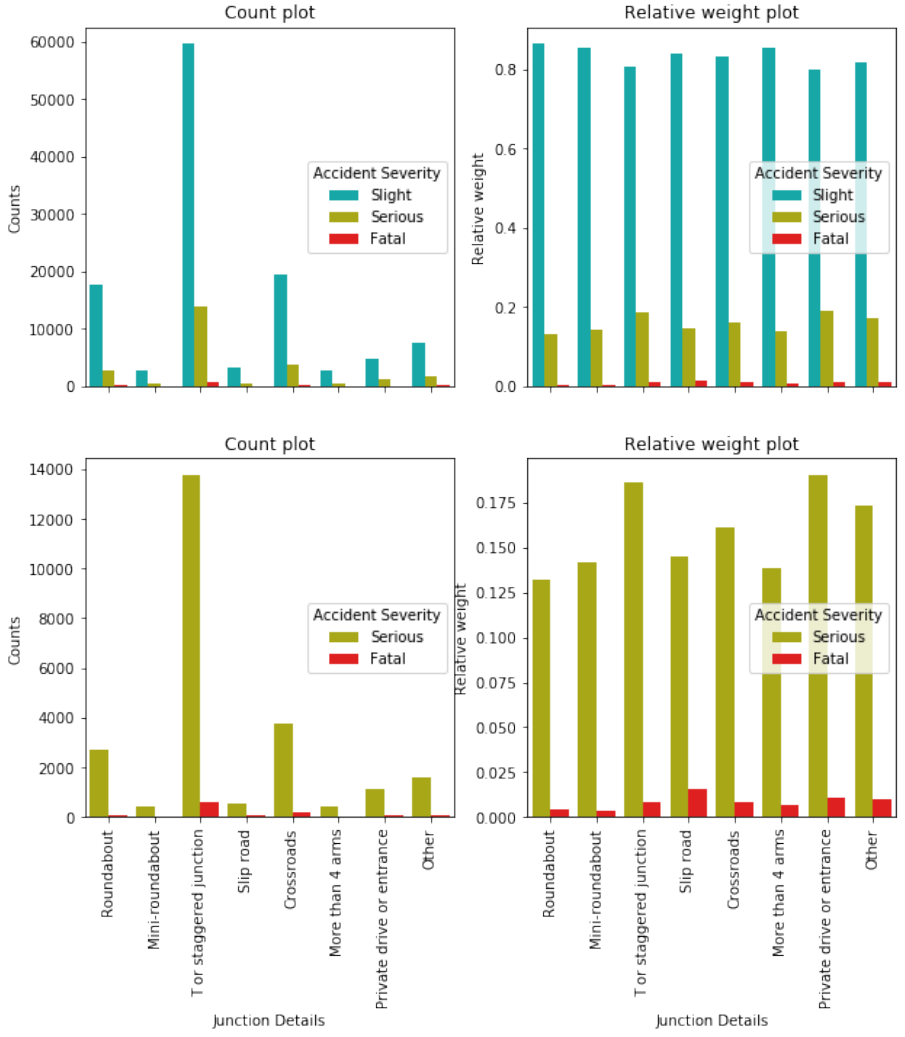
- **Most of the accidents** happened at **low speed limit** (30 mph), or 60 mph
- For **large speed** limit, the **relative weight** of **serious** and **fatal** accidents is **higher** (max at 60 mph)



- Left: Count distribution
- Right: Relative Weights (to counts at that particular x-axis value)

In the bottom row, only serious and fatal cases are shown

- **Most of the accidents** happened during **rush hours** (15:00-18:00)
- At **night** (23:00-5:00) the **relative weight** of **fatal** accidents is **larger** (max around 4:00)

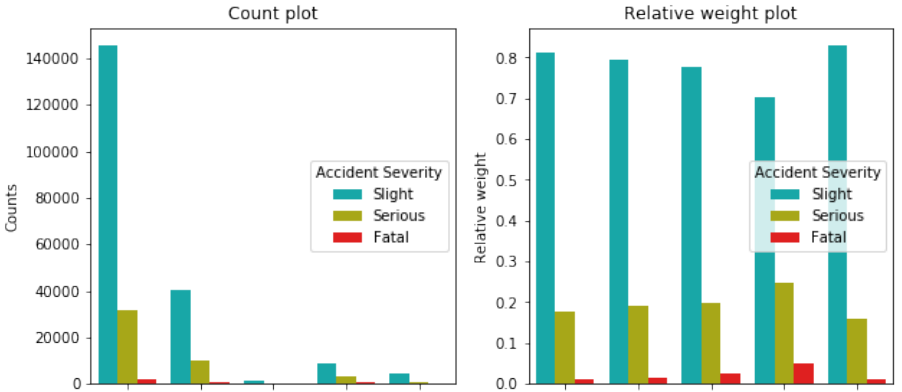


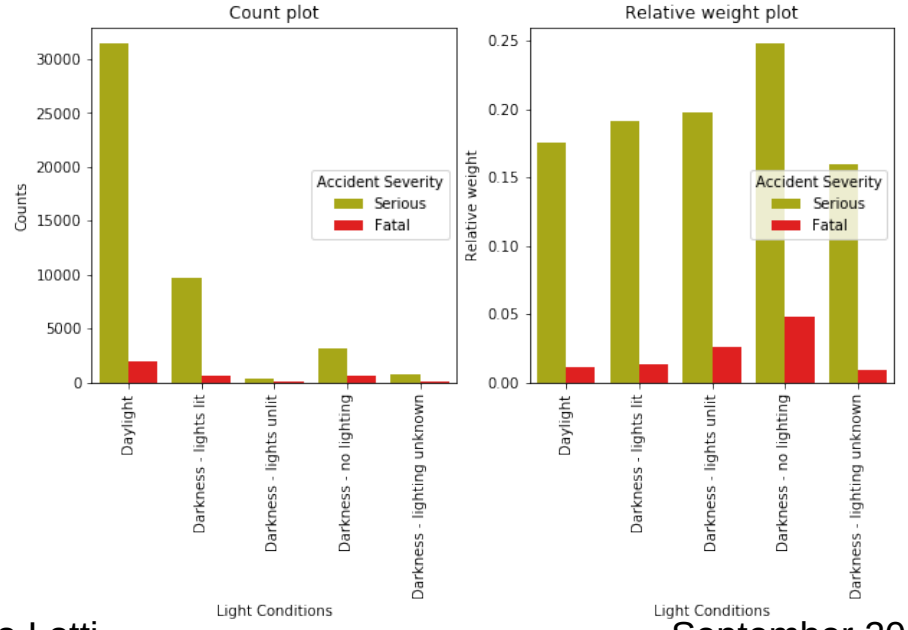
Junction Type

- **Most of the accidents** in correspondence of junctions happened at **T or staggered junctions**, **crossroads** and **roundabouts**
- **T or staggered junctions**, **private drives** and **slip road** junctions have the largest **relative weight** for **serious** and **fatal** accidents

- Left: Count distribution
- Right: Relative Weights (to counts at that particular x-axis value)

In the bottom row, only serious and fatal cases are shown



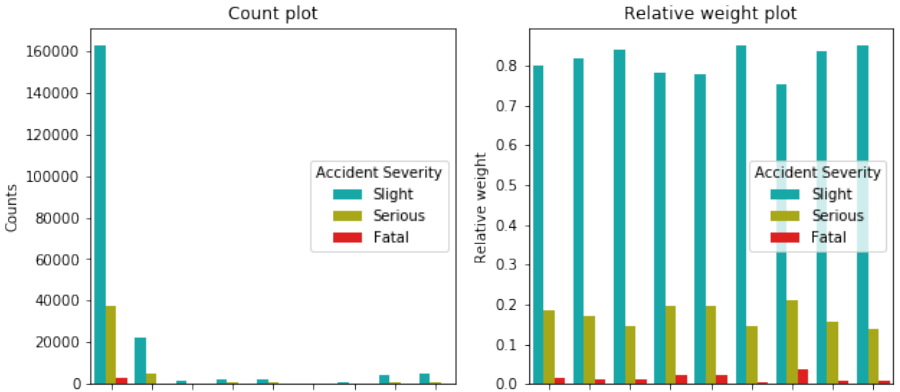


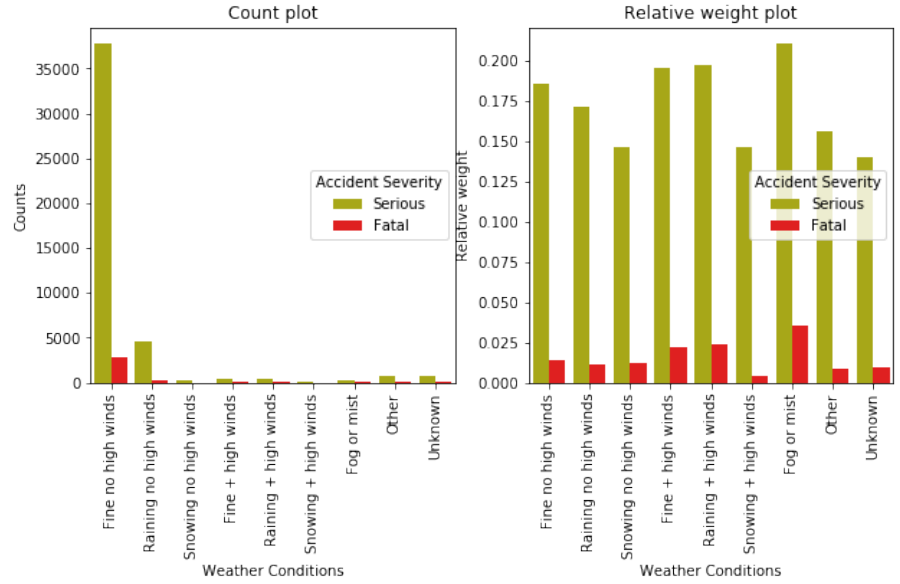
Light Conditions

- **Most of accidents** happened during **daylight**
- The **relative weight** of **serious** and **fatal** accidents is **higher for darkness** (max for darkness without lighting)

- Left: Count distribution
- Right: Relative Weights (to counts at that particular x-axis value)

In the bottom row, only serious and fatal cases are shown





Weather Conditions

- The **majority** of accidents happened for **fine** weather, followed by **rain**
- **Highest relative weights** for **fog** and **high wind** conditions

- Left: Count distribution
- Right: Relative Weights (to counts at that particular x-axis value)

In the bottom row, only serious and fatal cases are shown

The models

- **Classification models** were chosen
- The following models were considered:
  - **K-Nearest Neighbors**;
  - **Decision Tree**;
  - **Bootstrap Aggregation (Bagging) Classifier**;
  - **Adaptive Boosting (AdaBoost) Classifier**;
  - **Random Forest Classifier**.
- Hyper-parameter tuning was done using **balanced accuracy score**: average of recall (i.e. true positives divided by the number of real positives) obtained on each class

- **option class_weight='balanced'** was used: each sample contribution is weighted inversely proportional to its class frequency (i.e. total count of samples in that class). Less populated classes thus will count the same as the most populated ones. This is equivalent to **oversampling**.

  Such option was used for **Decision Tree** and **Random Forest** Classifiers, as well as the base estimators of **Bagging** and **AdaBoost** Classifiers.

- Two additional models using undersampling of the most populated classes were trained:

  - **Balanced Bagging Classifier**;

  - **Balanced Random Forest Classifier**.

Test the model

- The models were tested on the accident database from **2016**
- The data were prepared similarly to what done to the training dataset
- After data cleaning, 135815 samples (99.4% of initial data) were in the test dataset

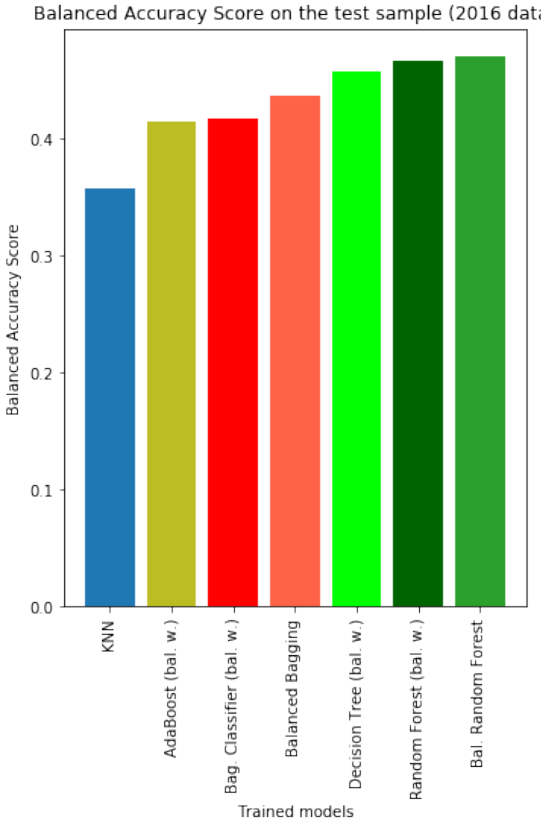
[accident and casualty adjustment 2004 to 2018]	ZIP	2019	Not available
Road Safety Data - Accidents 2017	ZIP	12 October 2018	Not available
Road Safety Data - Casualties 2017	ZIP	12 October 2018	Not available
Road Safety data -Vehicles 2017	ZIP	12 October 2018	Not available
Road Safety Data - Accidents 2016	ZIP/CSV	29 September 2017	Not available
Road Safety data -Vehicles 2016	ZIP/CSV	29 September 2017	Not available
Road Safety Data - Casualties 2016	ZIP/CSV	29 September 2017	Not available

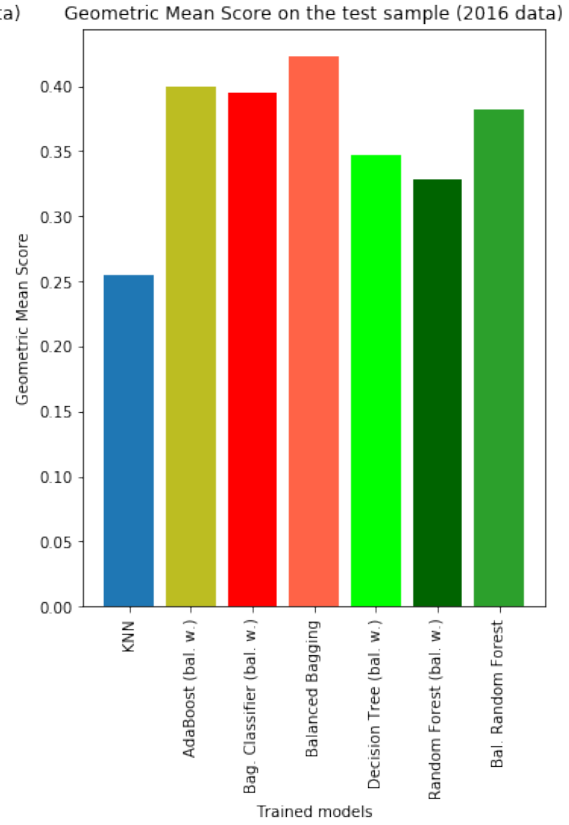
Results

Two **metrics** were considered:

- **Balanced Accuracy Score** from sklearn.metrics
- **Geometric Mean Score** from imblearn.metrics

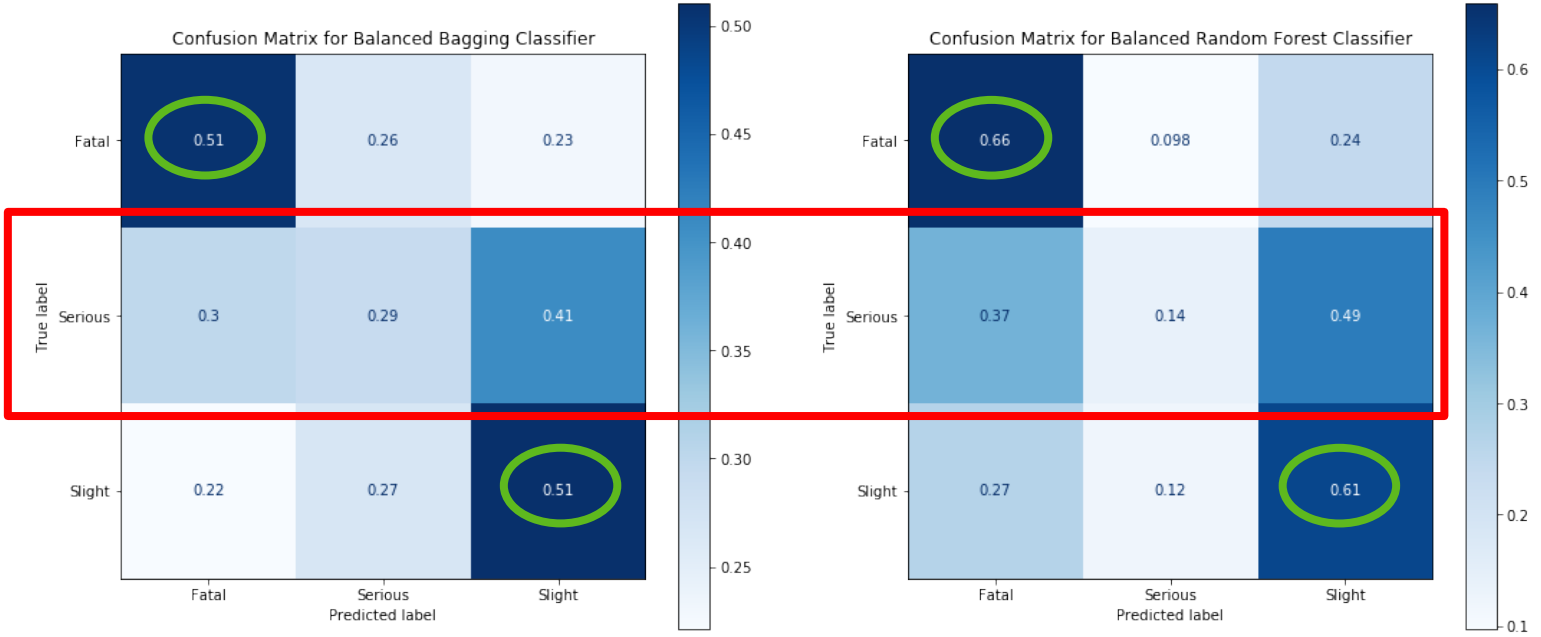
- Balanced Accuracy: average of recall scores (true positives correctly identified over the total number of true positives) per target class.
- Geometric Mean (G-Mean): Nth-root of the product of class recalls; N is the number of classes.
- For both: 0 worst possible value; 1 best possible value



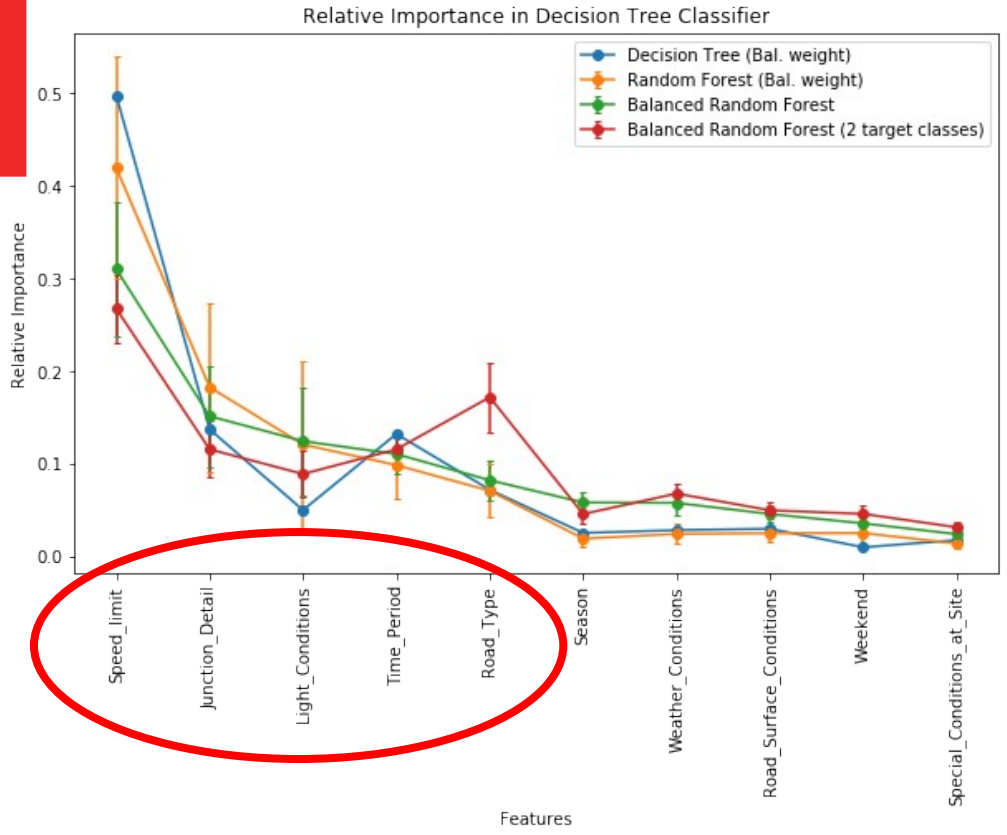


- **KNN** scored worst
- **Dec. Tree** and **Rand. Forest** models scored best in Bal. Acc. (0.46-0.47)
- **AdaBoost** and **Bagging models** scored best in G-Mean (0.40-0.42)
- **Balanced Random Forest** scored best in Bal. Acc. with 0.47
- **Balanced Bagging** scored best in G-Mean with 0.42

- **Ensemble models with data undersampling** performed best, according to these metrics
- All scores were not extremely high



- All models had **problems in correctly identifying serious accident cases**
- Bal. Bagging had the best performance in **serious** class, correct predictions ~**30%** of actual true cases; it reached recall around **51%** for both **fatal** and **slight** cases
- Bal. Random Forest had recall of **66%** for **fatal** and **61%** for **slight**. It still wrongly predicted ~49% of true serious cases as slight; recall for **serious** was ~**14%**



Relative Importance from trained models

- Most important feature: **Speed Limit**

  - relative weight of serious and fatal accid. increased with increasing speed limit

- Next important:

  - Junction Details

  - Light Conditions, Time Period

  - Road Type

- **Visibility** (Light Cond. - Time Period) is an important factor in the severity of accidents

- **Type of junction** is also a discriminant in severity prediction

Conclusions

- This study analyzed UK accidents data from 2017 + 2018
- It trained a series of models to predict the severity of accidents, concentrating on the ability to correctly predict critical accidents (fatal and serious)
- The best performances, among the models tested during the study, were reached by **undersampling ensemble models**: Balanced Bagging and Balanced Random Forest
- No model could achieve recall >50% for **all** the classes **at the same time**

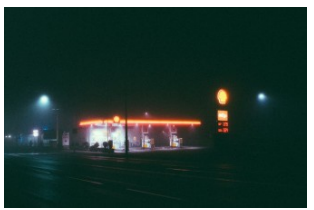
Conclusions – contd.

- It showed the relative importance of features from the trained models

Critical factors:

- **Speed Limit:** higher speed limits had relatively larger chance of high severity accidents

- **Visibility Conditions:** bad visibility (night, no lights) was a factor increasing the probability of high severity



- **Junction Types:** Roundabouts were relatively more safe T or staggered junctions and slip roads were the relatively most dangerous

Perspectives

Possible future developments:

- Generalization of the severity prediction based on road users (drivers, bikers, cyclists, pedestrians etc.) → target focused prevention measures; identify improved security devices for particularly dangerous conditions
- Use other classification models, such as Gradient Boosting, or include features which were dropped in this study
- Analysis of the evolution of risk factors during the years

- **Important note**: absolute counts (count histograms) are also very important, since the total cost for the society derives from a combination of **both relative weight** of severity, and **frequency of happening**.

Final Remarks

- This study was done in the context of the Coursera course: Applied Data Science Capstone of IBM Data Science Professional Certificate

  https://www.coursera.org/learn/applied-data-science-capstone

- The full notebook of the analysis can be found on Github:

  https://github.com/AntonioBL/Coursera_Capstone

All the photos in this presentations are free-to-use images downloaded from https://www.pexels.com