Severity of traffic road accidents in France

Prediction project applied for 2016 - 2017

Reasons and factors of high severity of traffic road accidents in France are multiple and partly unknown (in particular new causes)

How reduce the risk of accident and decrease severity level?
How collectivities can reinvent their actions to increase security?
How can each of us learn and reduce our high risk behavior?

Data acquisition and cleaning

- In France data.gouv.fr platform share data among others on road safety
- From 2016 to 2017, 120 000 accidents will be used to build a prediction model for the risk of high severity (death and serious injury)
- Incomplete or inconsistent data, correlated, not representative or useless for the prediction characteristics are deleted
- We keep 88 295 accidents and 19 features (153 after one-hot encoding)
- 3 features recreated for hour of the day, day of the week and 18 geographic areas instead of exact location
- Normalization of age of victims and date data (hour, day and month)

Many features in the accident conditions but almost no correlation

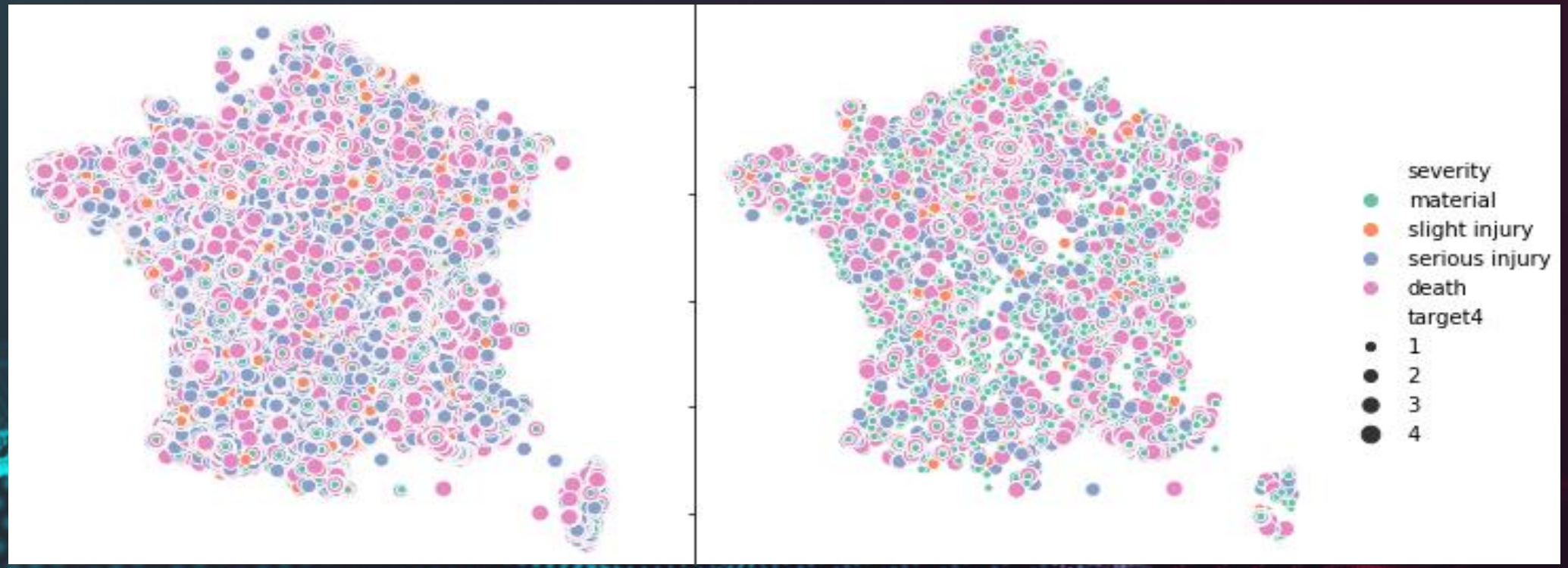
catu	grav	sexe	trajet	an_nais	senc	catv	occutc	obs	obsm	choc	manv	dayofweek	geo
0.006104	-0.001181	0.024018	0.080104	0.032102	0.049554	0.004624	0.014973	0.032224	0.013212	0.011618	0.032634	0.002839	-0.024260
0.006107	-0.001191	0.024029	0.080120	0.032093	0.049532	0.004603	0.014983	0.032262	0.013188	0.011608	0.032618	0.002856	-0.024100
-0.000078	-0.002592	0.022318	0.005393	0.007721	0.024228	0.000360	-0.025768	-0.012391	0.016194	0.006006	0.033743	-0.010321	0.002584
-0.002958	-0.000607	-0.001885	0.004111	0.001961	-0.005777	0.005745	0.000870	0.000174	0.005153	-0.001763	-0.000563	-0.008198	-0.007141
0.025379	0.035555	-0.049502	0.020843	0.124576	-0.000651	-0.020716	-0.009444	0.068908	-0.050684	-0.019093	-0.005886	0.098046	-0.039475
0.078473	0.001727	-0.000251	-0.022161	-0.005592	-0.154820	0.046917	-0.025397	-0.154084	0.046561	-0.067248	0.059462	-0.071598	-0.127543
-0.012335	-0.001277	-0.000966	-0.007625	-0.018511	-0.037347	0.027855	0.006749	-0.059655	0.060717	0.003983	0.073436	-0.017622	-0.071358
0.018279	0.000026	0.019515	-0.002057	-0.004727	0.022850	-0.027678	0.040180	0.038523	-0.036242	-0.000198	-0.002930	0.009171	-0.010191
0.283911	0.053263	0.022739	0.018134	-0.012540	-0.039921	0.015470	0.012897	0.315584	-0.441700	-0.022589	-0.066441	0.016842	-0.013890
0.015171	-0.013098	0.016236	0.031562	0.010308	-0.028482	-0.054751	0.021672	0.030399	-0.018480	-0.005426	-0.022078	0.018448	0.094767
0.013629	0.010399	-0.009622	-0.005276	0.031274	0.007241	-0.000937	-0.019358	0.004278	-0.011962	0.012093	0.024441	-0.000235	-0.391807
0.012753	-0.005063	-0.022571	-0.010404	0.001080	0.078161	0.039254	-0.002895	-0.030905	0.030132	0.008936	0.010488	-0.001793	-0.058611
0.000293	0.007877	-0.016474	0.004379	0.025332	0.024555	0.031655	-0.003351	-0.033365	0.023782	0.030030	0.036851	-0.014764	-0.507012

For example no correlation between professional vehicle involved and type of trip for business reason

Accidents and severity throughout the metropolitan area in 2016 and 2017

For private/personal trip

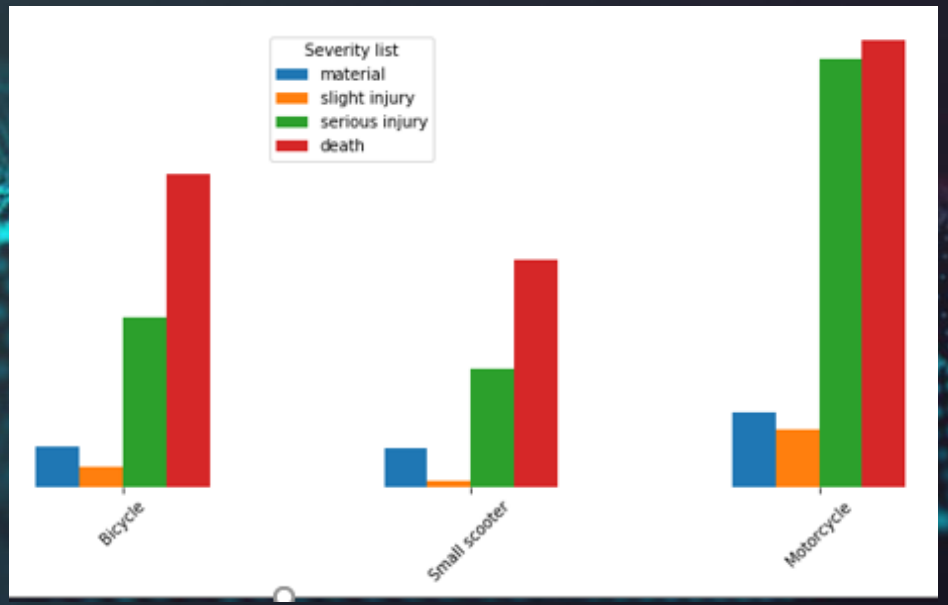
For business trip

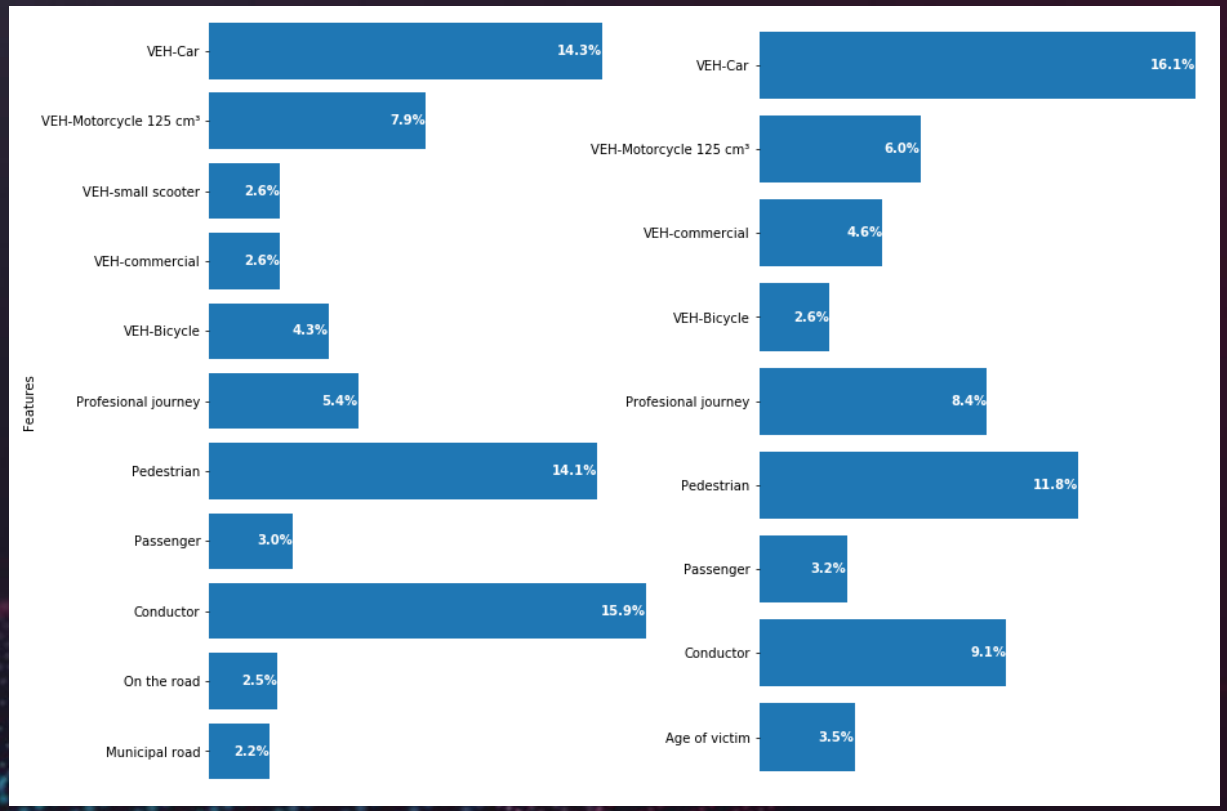


We can observe on the map of France accidents during business trips large white areas without accidents which correspond to forest areas or other solely natural or tourist areas. Use of commercial vehicle is another feature also associated with the least seriousness severity of accident

Classification models to explain results to stackholders and citizens show features of interest:

First group of vehicle use imports
But essentially bikes (bicycle, small scooter & motorcycle more than 125 cm³) receive more Severity – *moto have a different fatality distribution*





Random Forest vs X Gradient Boost forest

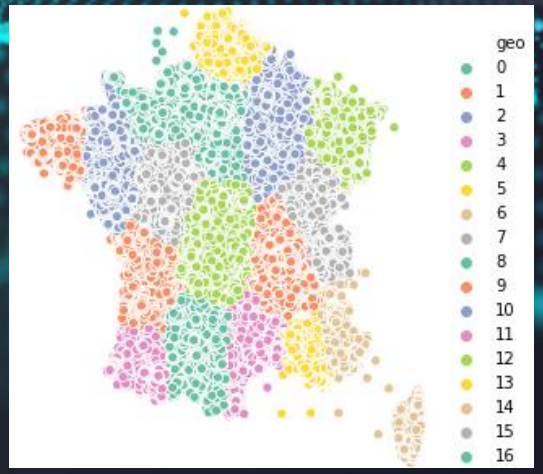
with the more important features determination

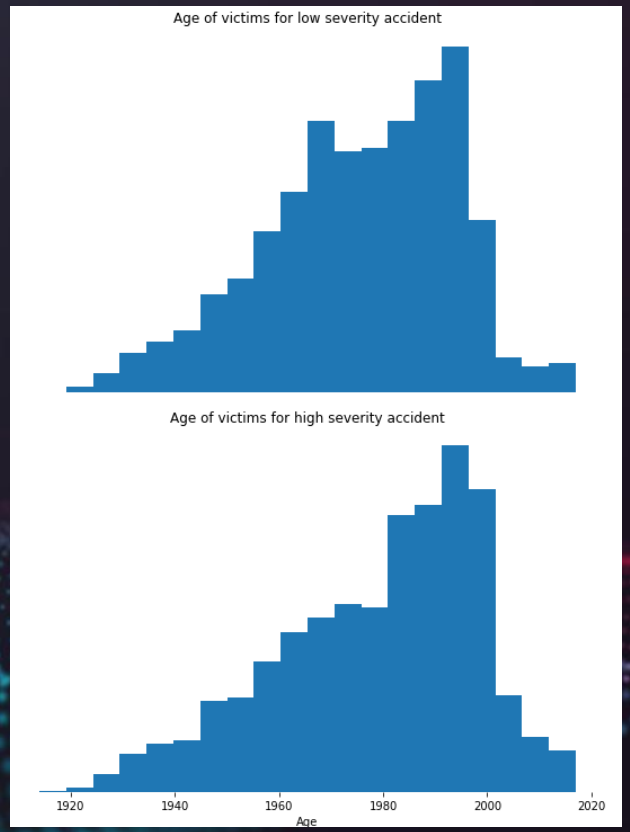
Record fatality ratio is for pedestrian

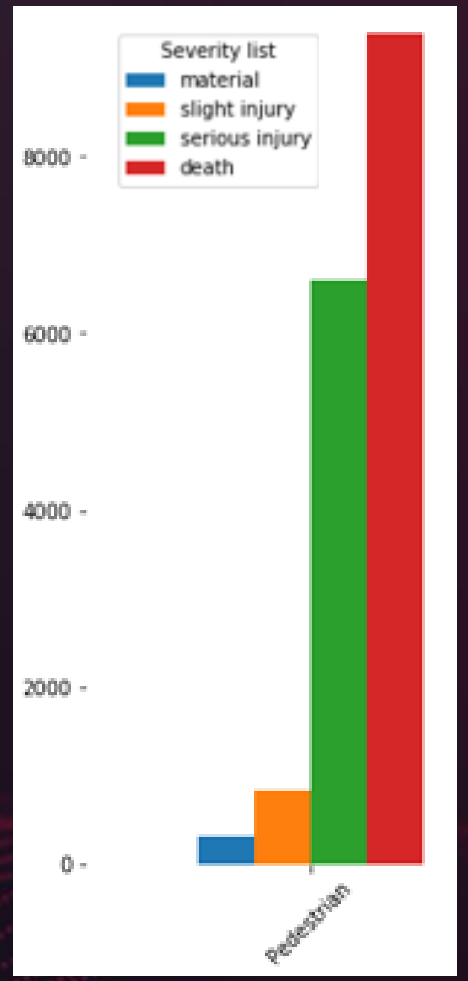
Number of death pedestrian give a mean of 1 for 10 accidents

Age of person is only relevant for increased risk of severity in the age group 15 – 40 years old

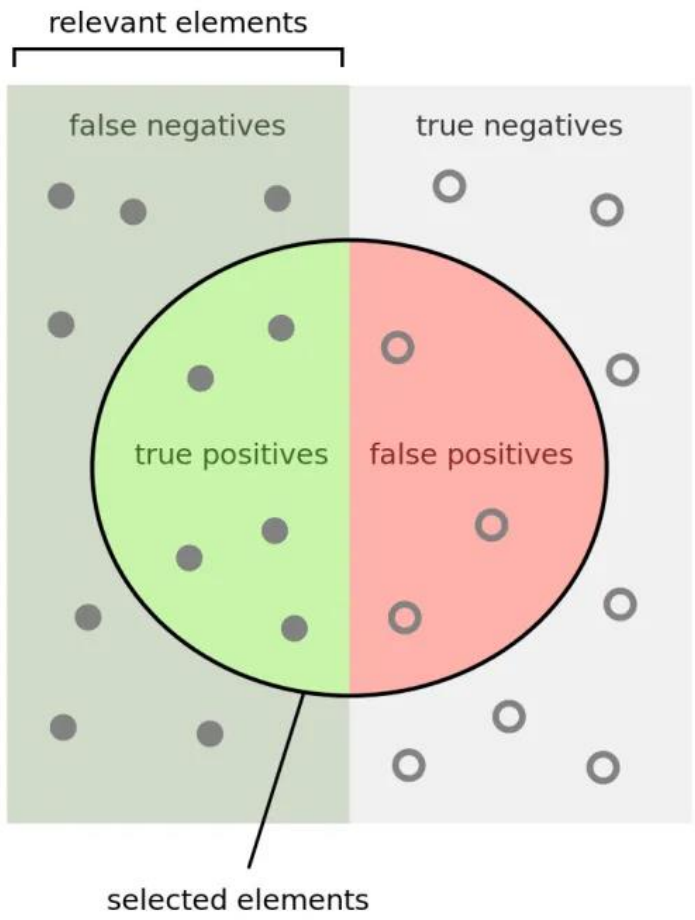
General geography doesn't reveal severity clustering







Analysis results



Final Model	Accuracy	Recall	F1-Score
Random forest	73 %	73 %	73 %
XGBoost forest	74 %	**74 %**	74 %

Best results for **XGBoost** for accuracy and F1-Score
but more relevant for the recall because we want to reduce false negative – a prediction of low severity need to limit mistakes in fatality risk

How many relevant items are selected?

$$\text{Recall} = \frac{\quad}{\quad}$$

F_1-score

$$\frac{1}{\frac{1}{2}\left(\frac{1}{\text{recall}} + \frac{1}{\text{precision}}\right)}$$

CONCLUSION

Strong impact of non-professional travel in increasing the risk of fatality.
We can work in awareness in induced risk taking when traveling in leisure or everyday transport with possible reduced attention.
To reduce the fatalities, we can work with ensemble of decisions trees that can be used for collectivities, awareness institutes and insurance companies.
Efforts in urban roads should also be increased compared to non-urban roads with less serious accidents.
Protection of persons with a better orientation for the pedestrians, and bicycle and motorcycle protection - Sensitize the population of the risk of these means of locomotion. For this, adjusting the educational discourse according to the locomotion behavior can be deepened.

And more reflexion

With the recent boom in small-capacity bicycle and motorcycle meal delivery businesses, it would be interesting to analyze this population separately, which could have the opposite effect with greater risk taking.

For even more relevant work of improve security, we need to include statistic information around this features like proportioning the number of use of the each mean of locomotion and depending of the other characteristics, like for instance, in urban locations like road without reserved cycle lane.