Predicting the severity of car accident

September 27, 2020 Vincent Hangard study project for training

1. Introduction

1.1 Background

To move for work, for see your family and friends, every day we use the road transports. But each use of the road is a risk of accident. Death or injury from traffic accident stay important each year and the effect of the ley begin to decrease strongly in France. Security of car is now better but if the death rate decreased, the accident with injury can be avoid too. Two measures continue to be strongly implemented, urban development for road safety, and awareness communication to citizens. Both need precision to have effect for save more lifes and have a big cost for the society. We are all concerned, the State must guarantee our safety and everyone plays a role in increasing or decreasing the risk of accidents.

1.2 Problem

Anyone can reduce the risk of an accident when we can predict the accident conditions that may occur, but we are not sufficiently educated. A particular accident zone is not sufficiently determined and the cause of the accident is sometimes unrecognized. Data that might contribute to determine the cause of accident with severity of personal injury or death needs to be better identified and a forecasting system could detect a risk area and condition. This project aims to predict severity risk and which condition (feature) will increase the risk.

1.3 Interest

The French road safety body needs a system to assess the risk of an accident based on geographic location and other criteria that most affect the risk of an accident. Insurance companies might be interested in using this model and features engineering to optimize thier coverage plans.

2. Data source and cleaning

2.1 Data sources

In France, data are shared in data.gouv.fr (FRANCE official open data).

https://www.data.gouv.fr/en/datasets/base-de-donnees-accidents-corporels-de-la-circulation/

Between 2016 and 2017, we are more than 120 000 accident cases.

Target prediction

There is characteristic information about accidents, locations, vehicles involved and victims. Accident severity is provided on 4 levels, which are unbalanced. Except that for our objective of preventing any accident with bodily impact, we can group the values 3 and 4 in serious severity (death and serious injury) and 1 and 2 in slight severity (material or sligthly) and data are correctly balanced.

2.2 Preprocessing

Rows with too much missing data are dropped or if a column have too much nan value, it's not considerated like feature and dropped too. Value of date of accident is split into day of the week (dayofweek column is created). Likewise, the encoding of the hour of the accident is a little less obvious than that of categorical variables. 24 corresponding categories will be considered at each hour of the day. For example, 00:20 corresponds to category 0. The pm 4:32 time corresponds to category 16 (if this feature will not appears important, it will be possible to double the spliting at 48 half-hours or more) - I replaced the hrmn column. 15 regions are created with a K-means clustering of Latitude and Longitude in 18 regions - column geo was created.

2.3 Features engineering

To optimize the training of the model, firstly, the correlation and variance matrices are calculated. The idea is that if variables are correlated, we can only keep one. This simplifies the data but can also make our model more robust since it reduces the impact of certain variables. Regarding variance, it is an indicator of the dispersion of the values of a certain variable. If it is low for a given variable, it will mean that the impact of this variable is negligible. Removing it allows us to gain in efficiency.

features = ['mois', 'lum', 'int', 'atm', 'catr', 'circ', 'nbv','vosp', 'prof', 'plan', 'surf', 'situ', 'catu', 'trajet', 'catv', 'dayofweek', 'geo'] in categories to encode in one-hot

and "an_nais" for age of person and "jour" for day of the month

2.4 Normalization

One-Hot encoding is chosen for categorical features with vectors composed of 0 everywhere and a 1 at the i -th modality using get_dummies from Pandas

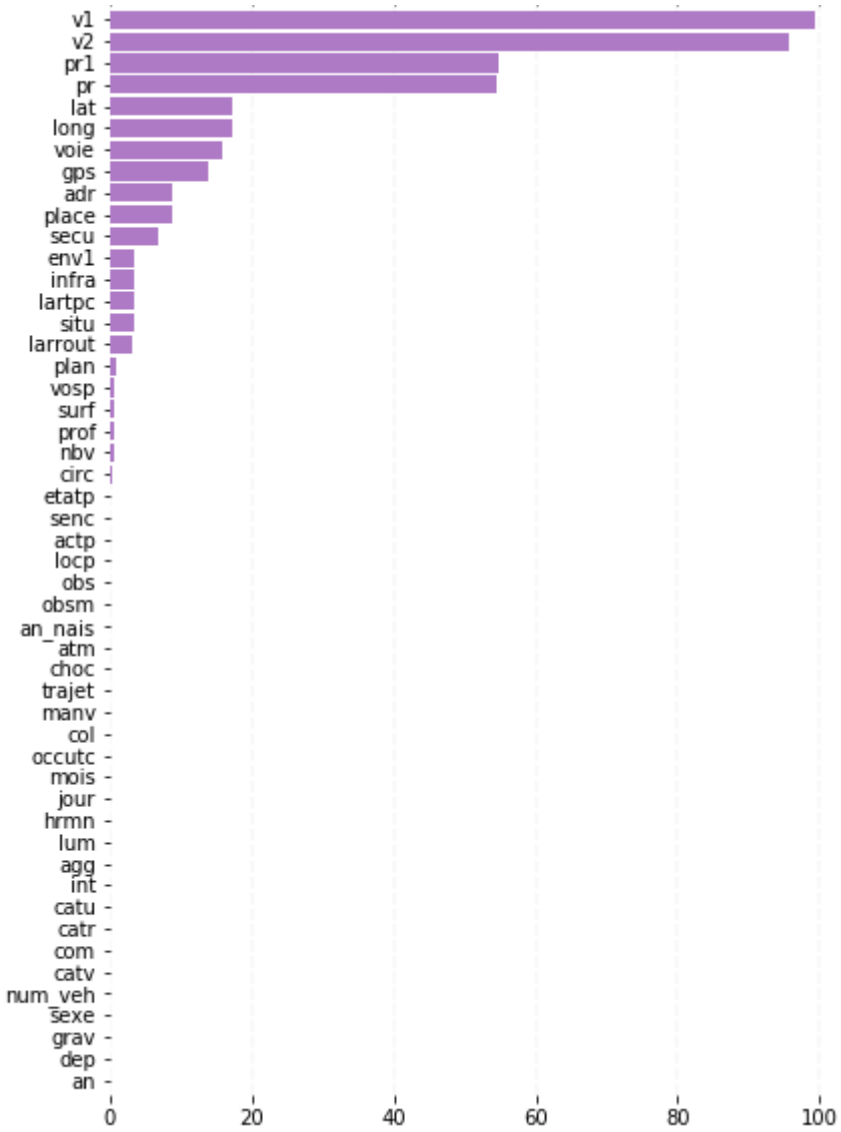
2.5 Last balance adjustement

Checking the distribution of the target value of severity, we adjust the training set to stay balanced.

3. Methodology

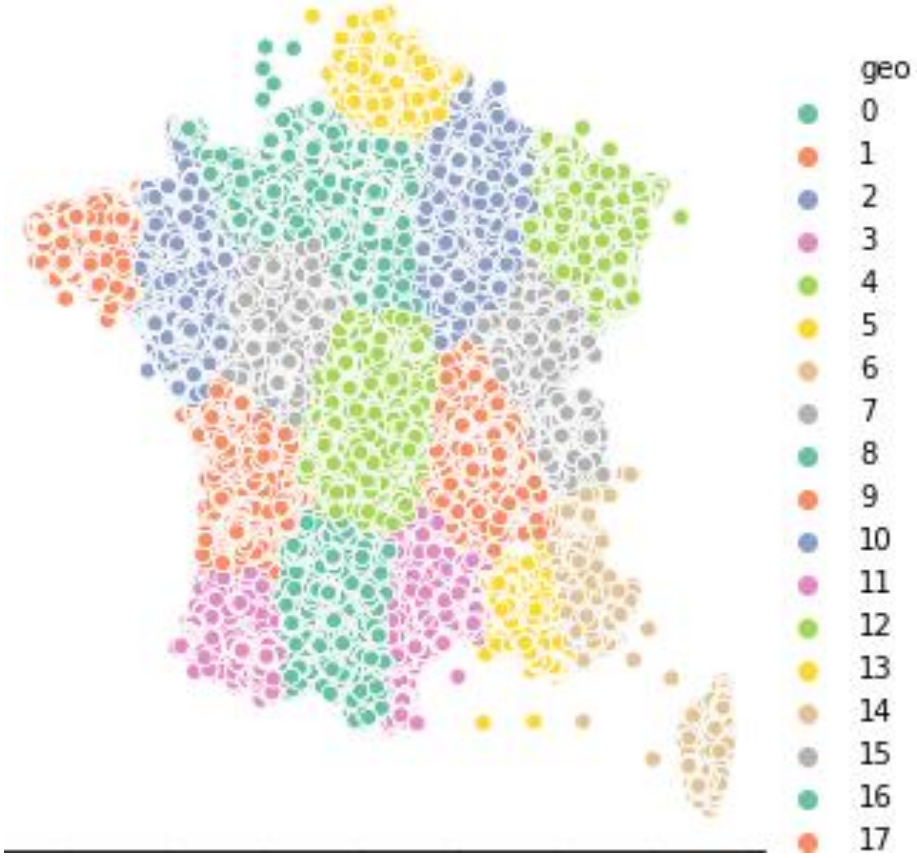
3.1 Data exploration

By looking at the quality of data intelligence, we retain the relevant data, here we have nan value count:



We keep only features with mor than 15% of completeness.

About the geographic splitting, it’s not visible that show a clustering but we can use it in the molelization in place of detailed localization: (Plotting with Longitude and latitude and the severity sizing the points – K-means method of clustering in Python)



Other observation about the potential correlation between the features that we kept: (plotting with heat map grid)

Between agg and catr, the category of road contains the agg binary possibility with a decline in 7 details of type of road.

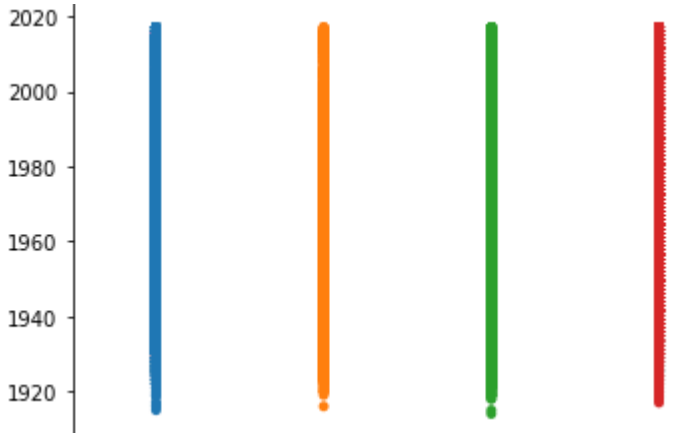
	Num_Acc	an	mois	jour	lum	agg	int	atm	col	com	lat	long	dep
Num_Acc	1.000000	1.000000	0.023712	0.006933	0.000511	-0.022939	-0.024221	0.017237	0.028695	0.003091	0.007856	0.051412	0.042014
an	1.000000	1.000000	0.023708	0.006932	0.000481	-0.023030	-0.024268	0.017252	0.028684	0.003118	0.007713	0.051430	0.041808
mois	0.023712	0.023708	1.000000	-0.004987	0.075086	0.004458	-0.003361	0.020956	-0.001027	0.007241	0.001966	-0.001409	0.008463
jour	0.006933	0.006932	-0.004987	1.000000	-0.009793	-0.000403	-0.008773	-0.022519	-0.007963	0.006123	0.000503	0.006330	0.000804
lum	0.000511	0.000481	0.075086	-0.009793	1.000000	0.114303	0.043569	0.028293	0.054550	-0.019494	0.041128	0.010886	0.040686
agg	-0.022939	-0.023030	0.004458	-0.000403	0.114303	1.000000	0.217760	-0.078531	0.052201	-0.110439	0.024489	0.063065	0.043080
int	-0.024221	-0.024268	-0.003361	-0.008773	0.043569	0.217760	1.000000	-0.011269	-0.036900	-0.053286	0.078237	-0.031040	0.043925
atm	0.017237	0.017252	0.020956	-0.022519	0.028293	-0.078531	-0.011269	1.000000	0.030386	0.029210	0.062234	-0.033849	0.023278
col	0.028695	0.028684	-0.001027	-0.007963	0.054550	0.052201	-0.036900	0.030386	1.000000	-0.001257	0.054187	-0.010465	0.026049
com	0.003091	0.003118	0.007241	0.006123	-0.019494	-0.110439	-0.053286	0.029210	-0.001257	1.000000	0.198357	-0.031688	0.040589
lat	0.007856	0.007713	0.001966	0.000503	0.041128	0.024489	0.078237	0.062234	0.054187	0.198357	1.000000	-0.269466	0.563701
long	0.051412	0.051430	-0.001409	0.006330	0.010886	0.063065	-0.031040	-0.033849	-0.010465	-0.031688	-0.269466	1.000000	-0.086971
dep	0.042014	0.041808	0.008463	0.000804	0.040686	0.043080	0.043925	0.023278	0.026049	0.040589	0.563701	-0.086971	1.000000
catr	-0.032785	-0.032799	-0.002431	-0.002735	0.046098	0.563847	0.168340	-0.037890	0.062838	-0.076388	-0.036611	-0.004149	-0.076323
circ	0.016885	0.016916	0.017243	0.003690	-0.007866	-0.229315	-0.124435	0.033549	-0.044866	0.090974	-0.001113	-0.044317	0.003312
nbv	0.043241	0.043177	0.005629	0.005049	0.023461	-0.170947	-0.079145	-0.001694	-0.045121	0.011357	0.076517	-0.018104	0.117430
vosp	-0.005234	-0.005256	0.006227	0.016172	0.015015	0.112910	0.025825	-0.008119	0.011225	-0.005369	0.024916	-0.006007	0.016575
prof	0.056782	0.056808	0.004191	0.005068	-0.003105	-0.094501	-0.046001	0.034155	-0.005471	0.005393	-0.033215	0.028284	-0.037720
plan	0.004661	0.004711	0.002164	-0.000580	-0.013085	-0.187231	-0.066931	0.048441	0.012181	0.036862	-0.060483	0.027865	-0.056813
surf	0.029025	0.029037	0.004323	-0.021770	0.060454	-0.101503	-0.036675	0.265951	0.063989	0.037437	0.058855	-0.025510	0.024845
situ	0.048841	0.048873	0.004704	0.002021	0.019502	-0.109948	-0.089821	0.028122	0.224674	0.038516	0.021809	-0.013562	-0.018576
catu	0.006104	0.006107	-0.000078	-0.002958	0.025379	0.078473	-0.012335	0.018279	0.283911	0.015171	0.013629	0.012753	0.000293
grav	-0.001181	-0.001191	-0.002592	-0.000607	0.035555	0.001727	-0.001277	0.000026	0.053263	-0.013098	0.010399	-0.005063	0.007877
sexe	0.024018	0.024029	0.022318	-0.001885	-0.049502	-0.000251	-0.000966	0.019515	0.022739	0.016236	-0.009622	-0.022571	-0.016474
trajet	0.080104	0.080120	0.005393	0.004111	0.020843	-0.022161	-0.007625	-0.002057	0.018134	0.031562	-0.005276	-0.010404	0.004379
an_nais	0.032102	0.032093	0.007721	0.001961	0.124576	-0.005592	-0.018511	-0.004727	-0.012540	0.010308	0.031274	0.001080	0.025332
senc	0.049554	0.049532	0.024228	-0.005777	-0.000651	-0.154820	-0.037347	0.022850	-0.039921	-0.028482	0.007241	0.078161	0.024555
catv	0.004624	0.004603	0.000360	0.005745	-0.020716	0.046917	0.027855	-0.027678	0.015470	-0.054751	-0.000937	0.039254	0.031655
occutc	0.014973	0.014983	-0.025768	0.000870	-0.009444	-0.025397	0.006749	0.040180	0.012897	0.021672	-0.019358	-0.002895	-0.003351
obs	0.032224	0.032262	-0.012391	0.000174	0.068908	-0.154084	-0.059655	0.038523	0.315584	0.030399	0.004278	-0.030905	-0.033365
obsm	0.013212	0.013188	0.016194	0.005153	-0.050684	0.046561	0.060717	-0.036242	-0.441700	-0.018480	-0.011962	0.030132	0.023782
choc	0.011618	0.011608	0.006006	-0.001763	-0.019093	-0.067248	0.003983	-0.000198	-0.022589	-0.005426	0.012093	0.008936	0.030030
manv	0.032634	0.032618	0.033743	-0.000563	-0.005886	0.059462	0.073436	-0.002930	-0.066441	-0.022078	0.024441	0.010488	0.036851
dayofweek	0.002839	0.002856	-0.010321	-0.008198	0.098046	-0.071598	-0.017622	0.009171	0.016842	0.018448	-0.000235	-0.001793	-0.014764
geo	-0.024260	-0.024100	0.002584	-0.007141	-0.039475	-0.127543	-0.071358	-0.010191	-0.013890	0.094767	-0.391807	-0.058611	-0.507012

	catr	circ	nbv	vosp	prof	plan	surf	situ	catu	grav	sexe	trajet	an_nais	senc	catv
Num_Acc	032785	0.016885	0.043241	-0.005234	0.056782	0.004661	0.029025	0.048841	0.006104	-0.001181	0.024018	0.080104	0.032102	0.049554	0.004624
an	032799	0.016916	0.043177	-0.005256	0.056808	0.004711	0.029037	0.048873	0.006107	-0.001191	0.024029	0.080120	0.032093	0.049532	0.004603
mois	002431	0.017243	0.005629	0.006227	0.004191	0.002164	0.004323	0.004704	-0.000078	-0.002592	0.022318	0.005393	0.007721	0.024228	0.000360
jour	002735	0.003690	0.005049	0.016172	0.005068	-0.000580	-0.021770	0.002021	-0.002958	-0.000607	-0.001885	0.004111	0.001961	-0.005777	0.005745
lum	046098	-0.007866	0.023461	0.015015	-0.003105	-0.013085	0.060454	0.019502	0.025379	0.035555	-0.049502	0.020843	0.124576	-0.000651	-0.020716
agg	563847	-0.229315	-0.170947	0.112910	-0.094501	-0.187231	-0.101503	-0.109948	0.078473	0.001727	-0.000251	-0.022161	-0.005592	-0.154820	0.046917
int	168340	-0.124435	-0.079145	0.025825	-0.046001	-0.066931	-0.036675	-0.089821	-0.012335	-0.001277	-0.000966	-0.007625	-0.018511	-0.037347	0.027855
atm	037890	0.033549	-0.001694	-0.008119	0.034155	0.048441	0.265951	0.028122	0.018279	0.000026	0.019515	-0.002057	-0.004727	0.022850	-0.027678
col	062838	-0.044866	-0.045121	0.011225	-0.005471	0.012181	0.063989	0.224674	0.283911	0.053263	0.022739	0.018134	-0.012540	-0.039921	0.015470
com	076388	0.090974	0.011357	-0.005369	0.005393	0.036862	0.037437	0.038516	0.015171	-0.013098	0.016236	0.031562	0.010308	-0.028482	-0.054751
lat	036611	-0.001113	0.076517	0.024916	-0.033215	-0.060483	0.058855	0.021809	0.013629	0.010399	-0.009622	-0.005276	0.031274	0.007241	-0.000937
long	004149	-0.044317	-0.018104	-0.006007	0.028284	0.027865	-0.025510	-0.013562	0.012753	-0.005063	-0.022571	-0.010404	0.001080	0.078161	0.039254
dep	076323	0.003312	0.117430	0.016575	-0.037720	-0.056813	0.024845	-0.018576	0.000293	0.007877	-0.016474	0.004379	0.025332	0.024555	0.031655
catr	000000	-0.322197	-0.353221	0.080824	-0.061615	-0.093685	-0.047476	-0.028590	0.074306	0.015131	0.003512	-0.012337	-0.009397	-0.246543	0.037979
circ	322197	1.000000	0.396590	0.001473	0.073851	0.061588	0.053528	0.063631	-0.034071	-0.014163	0.005224	0.056182	0.016897	0.099491	-0.024495
nbv	353221	0.396590	1.000000	0.074604	0.030806	-0.015149	0.019943	-0.014842	-0.059759	-0.006149	-0.017251	0.025884	0.020949	0.148722	0.012828
vosp	080824	0.001473	0.074604	1.000000	-0.018098	-0.045361	-0.014877	0.045131	0.018321	-0.002759	-0.005410	0.003745	-0.003620	-0.003410	0.033473
prof	061615	0.073851	0.030806	-0.018098	1.000000	0.238007	0.110081	0.062735	-0.007656	0.003994	0.003314	0.039085	0.010676	0.052344	-0.002707
plan	093685	0.061588	-0.015149	-0.045361	0.238007	1.000000	0.131646	0.119331	-0.020136	0.026183	-0.000954	0.047469	0.026248	0.051876	0.011707
surf	047476	0.053528	0.019943	-0.014877	0.110081	0.131646	1.000000	0.066232	0.002800	0.024569	0.018238	0.007231	0.020861	0.028435	-0.007561
situ	028590	0.063631	-0.014842	0.045131	0.062735	0.119331	0.066232	1.000000	0.019955	0.044308	-0.003030	0.053603	0.026574	0.023199	-0.035371
catu	074306	-0.034071	-0.059759	0.018321	-0.007656	-0.020136	0.002800	0.019955	1.000000	0.252255	0.209671	-0.014655	0.056389	-0.024059	-0.046833
grav	015131	-0.014163	-0.006149	-0.002759	0.003994	0.026183	0.024569	0.044308	0.252255	1.000000	0.096524	-0.014444	0.089471	0.003107	0.162994
sexe	003512	0.005224	-0.017251	-0.005410	0.003314	-0.000954	0.018238	-0.003030	0.209671	0.096524	1.000000	-0.011875	-0.042293	-0.004454	-0.138865
trajet	012337	0.056182	0.025884	0.003745	0.039085	0.047469	0.007231	0.053603	-0.014655	-0.014444	-0.011875	1.000000	-0.043742	0.042421	-0.019384
an_nais	009397	0.016897	0.020949	-0.003620	0.010676	0.026248	0.020861	0.026574	0.056389	0.089471	-0.042293	-0.043742	1.000000	0.000547	0.047069
senc	246543	0.099491	0.148722	-0.003410	0.052344	0.051876	0.028435	0.023199	-0.024059	0.003107	-0.004454	0.042421	0.000547	1.000000	0.024690
catv	037979	-0.024495	0.012828	0.033473	-0.002707	0.011707	-0.007561	-0.035371	-0.046833	0.162994	-0.138865	-0.019384	0.047069	0.024690	1.000000
occutc	008723	0.019955	0.034082	0.042501	0.013030	0.022433	0.018002	0.009925	0.048064	-0.025873	0.017103	0.005978	0.029429	0.010933	0.159178
obs	057334	0.042294	-0.013646	-0.027582	0.047485	0.137039	0.090099	0.323949	-0.009541	0.108700	-0.010389	0.048902	0.056152	0.008023	0.013963
obsm	001982	0.003190	0.041414	0.016224	-0.013444	-0.073833	-0.071901	-0.203812	-0.157066	-0.099874	-0.010070	-0.023687	-0.018996	0.028126	-0.007328
choc	079263	0.039374	0.072343	-0.007765	0.015871	0.018737	0.014688	0.014019	-0.064600	-0.031794	0.017195	0.026748	-0.006912	0.043386	-0.025612
manv	042709	-0.006367	-0.014560	0.021129	0.005567	-0.016052	-0.021662	0.022277	-0.055400	-0.075764	0.000294	0.031649	-0.003137	0.016959	-0.021933
dayofweek	037459	0.016935	-0.001057	-0.019517	0.018931	0.044696	0.025094	0.047644	0.024802	0.012620	-0.005041	0.080293	0.055512	0.009244	-0.028093
geo	017511	0.088581	-0.053436	-0.054710	0.030252	0.089084	0.002977	0.039810	0.006458	-0.029766	0.026472	0.053842	-0.025282	-0.010713	-0.052886

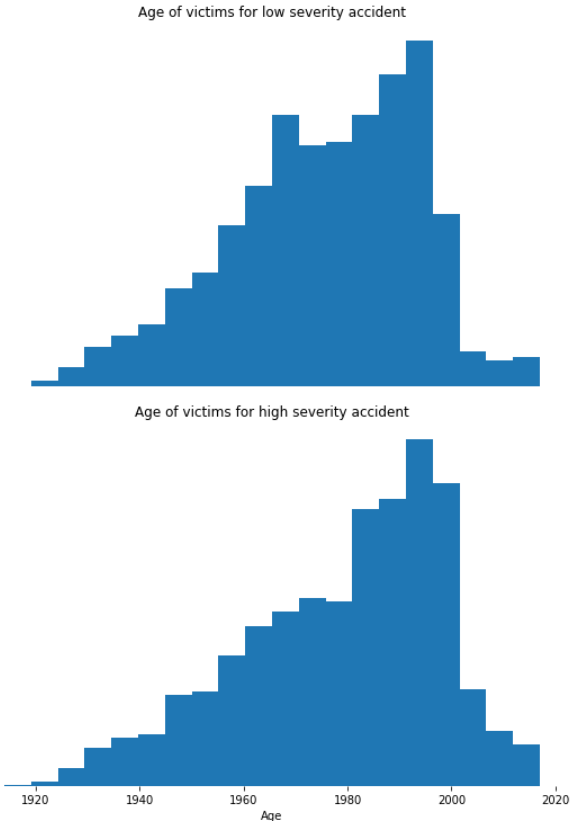
	occutc	obs	obsm	choc	manv	dayofweek	geo
Num_Acc	0.014973	0.032224	0.013212	0.011618	0.032634	0.002839	-0.024260
an	0.014983	0.032262	0.013188	0.011608	0.032618	0.002856	-0.024100
mois	-0.025768	-0.012391	0.016194	0.006006	0.033743	-0.010321	0.002584
jour	0.000870	0.000174	0.005153	-0.001763	-0.000563	-0.008198	-0.007141
lum	-0.009444	0.068908	-0.050684	-0.019093	-0.005886	0.098046	-0.039475
agg	-0.025397	-0.154084	0.046561	-0.067248	0.059462	-0.071598	-0.127543
int	0.006749	-0.059655	0.060717	0.003983	0.073436	-0.017622	-0.071358
atm	0.040180	0.038523	-0.036242	-0.000198	-0.002930	0.009171	-0.010191
col	0.012897	0.315584	-0.441700	-0.022589	-0.066441	0.016842	-0.013890
com	0.021672	0.030399	-0.018480	-0.005426	-0.022078	0.018448	0.094767
lat	-0.019358	0.004278	-0.011962	0.012093	0.024441	-0.000235	-0.391807
long	-0.002895	-0.030905	0.030132	0.008936	0.010488	-0.001793	-0.058611
dep	-0.003351	-0.033365	0.023782	0.030030	0.036851	-0.014764	-0.507012
catr	-0.008723	-0.057334	-0.001982	-0.079263	0.042709	-0.037459	-0.017511
circ	0.019955	0.042294	0.003190	0.039374	-0.006367	0.016935	0.088581
nbv	0.034082	-0.013646	0.041414	0.072343	-0.014560	-0.001057	-0.053436
vosp	0.042501	-0.027582	0.016224	-0.007765	0.021129	-0.019517	-0.054710
prof	0.013030	0.047485	-0.013444	0.015871	0.005567	0.018931	0.030252
plan	0.022433	0.137039	-0.073833	0.018737	-0.016052	0.044696	0.089084
surf	0.018002	0.090099	-0.071901	0.014688	-0.021662	0.025094	0.002977
situ	0.009925	0.323949	-0.203812	0.014019	0.022277	0.047644	0.039810
catu	0.048064	-0.009541	-0.157066	-0.064600	-0.055400	0.024802	0.006458
grav	-0.025873	0.108700	-0.099874	-0.031794	-0.075764	0.012620	-0.029766
sexe	0.017103	-0.010389	-0.010070	0.017195	0.000294	-0.005041	0.026472
trajet	0.005978	0.048902	-0.023687	0.026748	0.031649	0.080293	0.053842
an_nais	0.029429	0.056152	-0.018996	-0.006912	-0.003137	0.055512	-0.025282
senc	0.010933	0.008023	0.028126	0.043386	0.016959	0.009244	-0.010713
catv	0.159178	0.013963	-0.007328	-0.025612	-0.021933	-0.028093	-0.052886
occutc	1.000000	0.014841	-0.009712	0.004977	-0.011799	-0.009801	0.030053
obs	0.014841	1.000000	-0.337114	0.019127	0.000994	0.081105	0.046337
obsm	-0.009712	-0.337114	1.000000	0.037315	0.053550	-0.040960	-0.043971
choc	0.004977	0.019127	0.037315	1.000000	0.130489	0.002013	-0.022767
manv	-0.011799	0.000994	0.053550	0.130489	1.000000	-0.014217	-0.037887
dayofweek	-0.009801	0.081105	-0.040960	0.002013	-0.014217	1.000000	0.020680
geo	0.030053	0.046337	-0.043971	-0.022767	-0.037887	0.020680	1.000000

All features seems give possibility to use for predict and identify sensibility to the severity.

About age of people in the accident, we can see that everybody es exponed to accident for all severity:



And the real proportion of age reveals a slight disparity for the more old population which have more probability of minor accident than serious accident (between 15 and 40 old year):

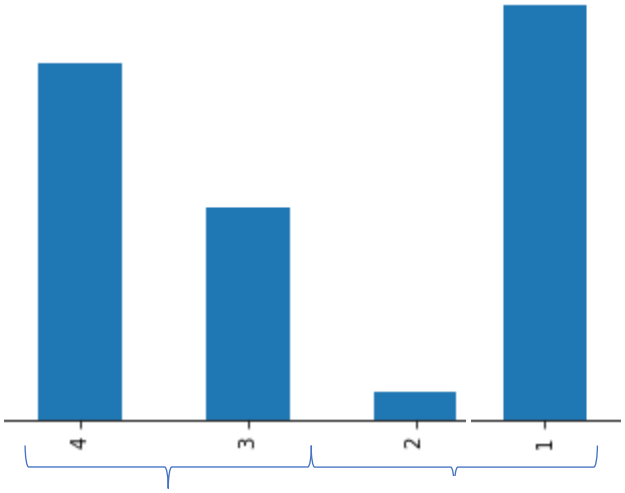


We can keep the 33 features corresponding of 4 groups:

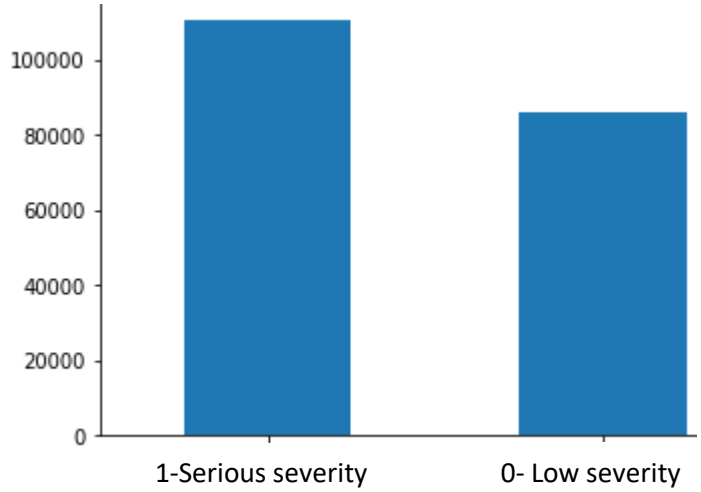
- Accidents conditions
- Accidents locations
- Vehicles involved
- Victims caracteristics

3.2 Model choice

A model using decision trees can be use for working with categoricals implications without lost a possibility to communicate about the understanding of the risk causes. And using ensemble of ramdom forest or X gradient boost forest to optimise the accuracy and prevent a bad recall. The question of severity with risk of death implicate we need to reduce the false negative and increase the true positive in serious severity.



And we have a correct balancing to the serious severity prediction:



3.2 Results

Ramdom Forest

With a grid search, we can found a Good balance with 100 estimators and a Depth of 8 or 9.

```
: # Normalization :
X = normalize(X.values)

X_train_rf, X_test_rf, y_train_rf, y_test_rf = train_test_split(X,y)

model_rf = RandomForestClassifier(n_estimators=100,
                                  max_depth=9
)

model_rf.fit(X_train_rf, y_train_rf)

# Test predictions and training data to compare

predictions_test = model_rf.predict(X_test_rf)

predictions_train = model_rf.predict(X_train_rf)

# Accuracy of both

train_acc = accuracy_score(y_train_rf, predictions_train)
print(train_acc)

test_acc = accuracy_score(y_test_rf, predictions_test)
print(test_acc)
```

```
0.7333115828252551
0.7287223196443865
```

72.8 % for the test data

XGradient Boost

With a grid search, we can found a Good balance with 100 estimators and a Depth of 8 or 9.

```
]: X_train, X_test, y_train, y_test = train_test_split(X,y)

model_boosting = GradientBoostingClassifier(loss="deviance",
    learning_rate=0.25,
    max_depth=5,
    max_features="sqrt",
    subsample=0.95,
    n_estimators=200)


model_boosting.fit(X_train, y_train)

# On calcul les prédictions
predictions_test_xgb = model_boosting.predict(X_test)
predictions_train_xgb = model_boosting.predict(X_train)

# On affiche les résultats :

train_acc = accuracy_score(y_train, predictions_train_xgb)
print(train_acc)

test_acc = accuracy_score(y_test, predictions_test_xgb)
print(test_acc)
```

```
0.7547290362485811
0.7392031320092981
```

We obtain a 73.9 % of accuracy, a little better than ramdom forest

Comparing Random Forest and XGB

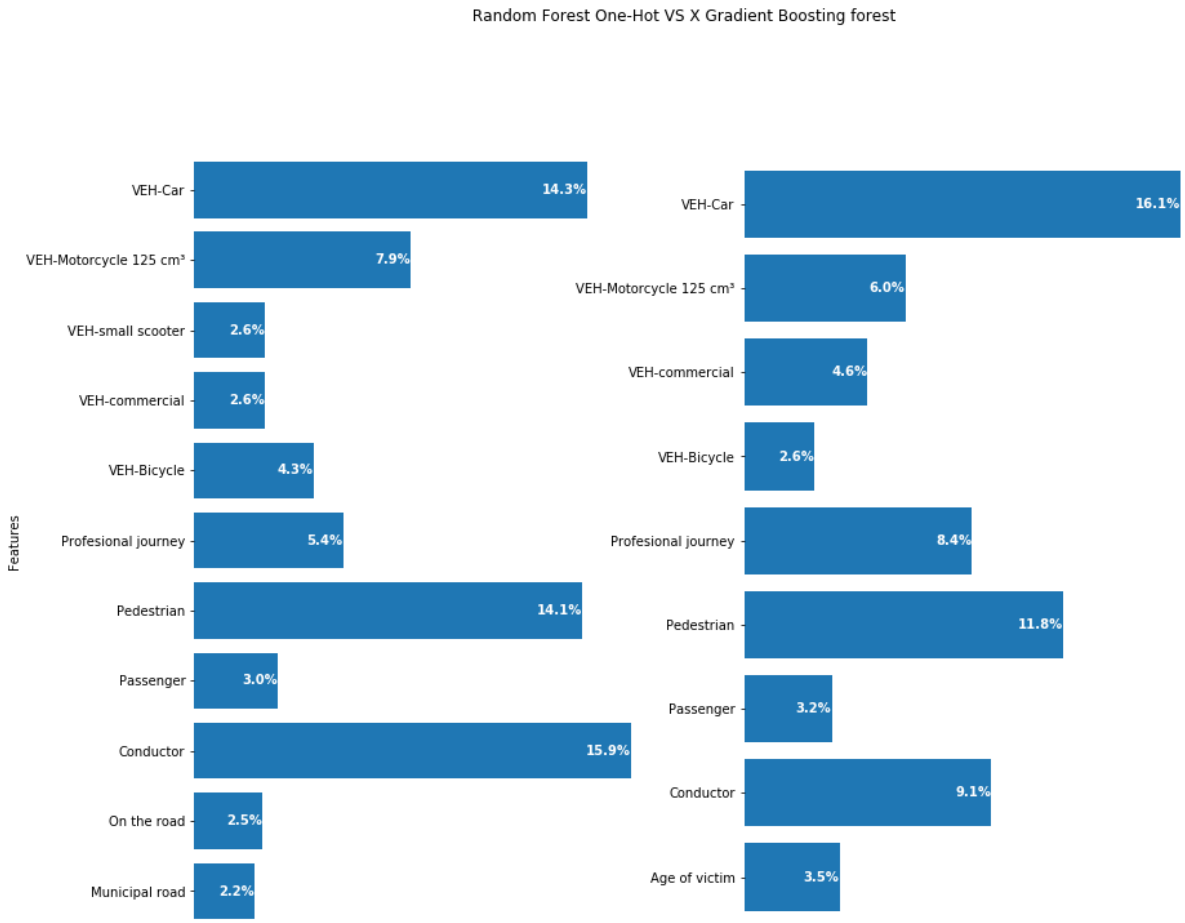
```
]: # Calcul du recall pour Random Forest

recall_rf = recall_score(y_test_rf, predictions_test, average='macro')
print('Recall: %.3f' % recall_rf)

# Calcul du recall pour XGBoost

recall = recall_score(y_test, predictions_test_xgb, average='macro')
print('Recall: %.3f' % recall)

# Calcul du F1-Score pour Random Forest

f1_rf = f1_score(y_test_rf, predictions_test, average='macro')
print('F1-Score: %.3f' % f1_rf)

# Calcul du F1-Score pour XGBoost

f1 = f1_score(y_test, predictions_test_xgb, average='macro')
print('F1-Score: %.3f' % f1)
```

```
Recall: 0.728
Recall: 0.739
F1-Score: 0.726
F1-Score: 0.737
```

Recall are correct 72.8% for random forest and 73.9% for XGB.

3.3 Discussion

Both are working correctly and give an interesting diversity of importance in the features:

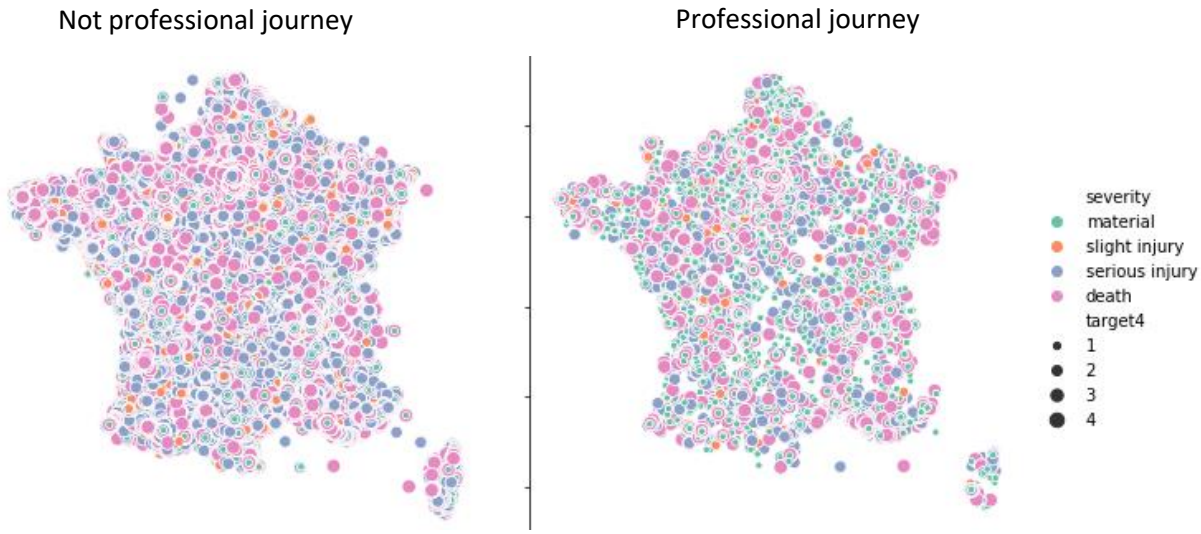


For first group of features we can see in this chart, Vehicle involved confirm other statistict of the French traffic road institute of prevention, Car is more involved in the risk of severity, but commercial vehicle for material accident. And the small vehicle like bicycle (or walking like pedestrian) which are locomotion with low security equipment increase the risk of severity for its driver. Motorcycle for the speed capacity to weight and limited security ratio increase fatality risk.

Municipal road type is really indicating a bigger risk of severity than other roads.

One observation interesting concerns the 2 features of professional journey and commercial vehicle which are decisive for a low severity. We can see that a traject for professional reason reduce the severity risk.

For accident characteristic, one type of trip is important but in the reverse implication, professional journey is more sure than all other type of road trip:



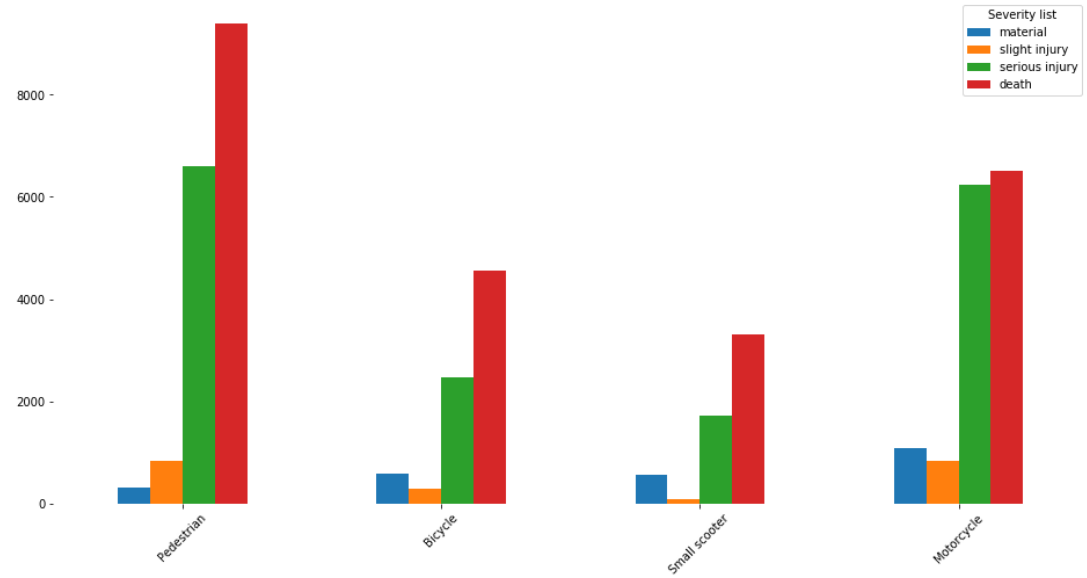
We can observe on the map of France accidents during business trips (right map), large white areas without accidents which correspond to forest areas or other solely natural or tourist areas.

For victim feature group, Pedestrian and Conductor are principals victims

And for the infrastructure incidence, municipal road show more severity of accident and the real location is on the road and not out of the road.

The last indicator for the increasing fatality risk is the use of 4 type of tansport:

- Walking is the means of locomotion with highest aggravationg risk
- By Motorcycle (more than 125 cm^3)
- By small scooter
- By bicycle



4. Conclusion

In France, in 2016 and 2017 we know that almost half of the accidents resulted in fatalities and serious injuries.

We have also discovered the strong impact of non-professional travel reasons in increasing the risk of fatality with two features (reason of journey and profesional vehicle used). We can work in awareness ininduced risk taking when we traveling in leisure or everyday transporte with possible reduced attention.
On the other hand, 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.

But to find new improvements, to reduce the fatality, we can work in 3 areas with which ML model of ensemble of decision tree can be used for collectivities, awareness institutes and insurance company.

Among the axes, we have the urban development oriented in the type of road and to target the protection the road itself, more than the rest of urban area. Efforts on urban roads should also be increased compared to non-urban roads with less serious accidents. Then the protection of person with a better orientation for the pedestrian, and bicycle and motorcycle protection (It’s visible than moto protection distributes mortality by reduction to severe injuries in compartion to bikes and light scooter). Sensitize the population of the risk of these means of locomotion. For this adjusting the educational discourse according to the locomotion behavior and according to the rol like the driver.

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 example, in a type of urban location like road without reserved cycle lane.