1 Tree based configuration

Table 1 presents the configuration of tree based models.

Table 1: **Configuration of each tree based model**

Parameters	**XGBoost**	**Catboost**
Early Stopping Rounds	15	15
Learning Rate	0.001	0.001
Max Depth/Num Leaves	6	4
Min Child Weight	1.0	-
Max Delta Step	1.0	-
Subsample	0.5	0.7
Col Sample Bytree	0.7	0.6
Reg Lambda	1.7	1
Reg Alpha	0.7	0.27
N Estimators	10000	10000
Tree Method	hist	-

2 Deep learning layers configuration

This section present the layers parameters for the MLP model used in this study.

2.1 MLP

Table 2 shows the configuration of the MLP models.

Table 2: MLP configuration

Linear 1	Linear 2	Linear 3	Linear 4
179, ReLU	179, ReLU	64, ReLU	5

2.2 GRU

Table 3 shows the configuration of the GRU model.

Table 3: GRU configuration

GRU	Norm 1	Dropout	Linear 1	Linear 2	Linear 3
2 layer, 0.03 dropout, 179, ReLU	BatchNorm	0.03	256, ReLU	64, ReLU	5

3 Results

3.1 Classic evaluation

Figure 1 shows the F1 score achieved by each model on the full dataset.

3.2 Generalization evaluation

Figure 2 shows the generalisation F1 score achieved by each model.

Figure 3 plots how the generalization score evolves for three approaches—(i) the standalone CatBoost model, (ii) CatBoost trained with SMOTE (oversampling factor 6), and (iii) a 10-model voting ensemble—while progressively adding departments sorted in ascending order by the number of fires recorded in 2023.

Figure 4 plots how the generalization score evolves for three approaches—(i) the standalone Xgboost model, (ii) CatBoost trained with SMOTE (oversampling factor 6), and (iii) a 11-model voting ensemble—while progressively adding departments sorted in ascending order by the number of fires recorded in 2023.

Figure 5 plots how the generalization score evolves for three approaches—(i) the standalone Xgboost model, (ii) CatBoost trained with SMOTE (oversampling factor 6), and (iii) a 11-model voting ensemble—while progressively adding departments sorted in ascending order by the number of fires recorded in 2023.

Figure 6 plots how the generalization score evolves for three approaches—(i) the standalone LG model, (ii) CatBoost trained with SMOTE (oversampling factor 6), and (iii) a 12-model voting ensemble—while progressively adding departments sorted in ascending order by the number of fires recorded in 2023.

Figure 7 plots how the generalization score evolves for three approaches—(i) the standalone LG model, (ii) CatBoost trained with SMOTE (oversampling factor 6), and (iii) a 12-model voting ensemble—while progressively adding departments sorted in ascending order by the number of fires recorded in 2023.

3.3 Extreme events evaluation

Figure 8 shows the F1 score achieved by each model on samples (predicted or true) of class 2, 3 or 4. Figure 9 shows the F1 score achieved by each model on samples (predicted or true) of class 3 or 4.

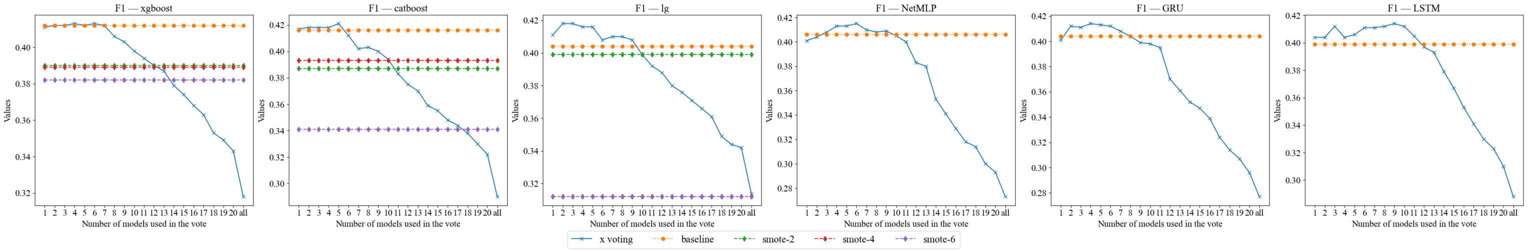

Figure 1: F1 performance on samples of true risk (or predicted) superior or equal than 2 between voting, classic and SMOTE models.

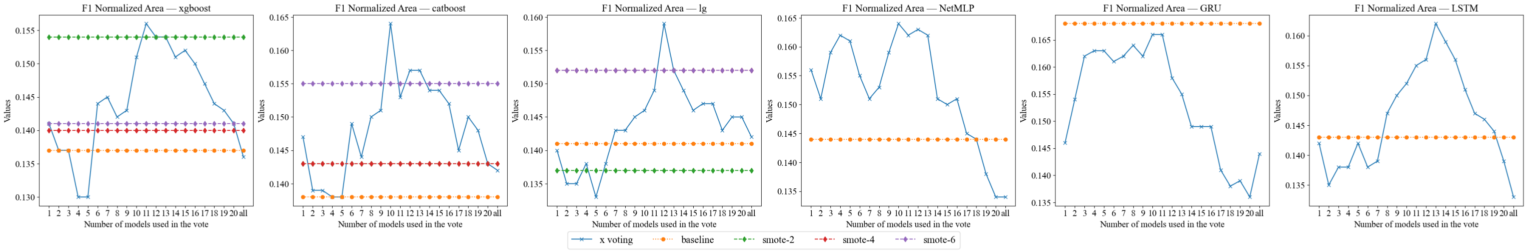

Figure 2: Area F1 performance on full database between voting, classic and SMOTE models.

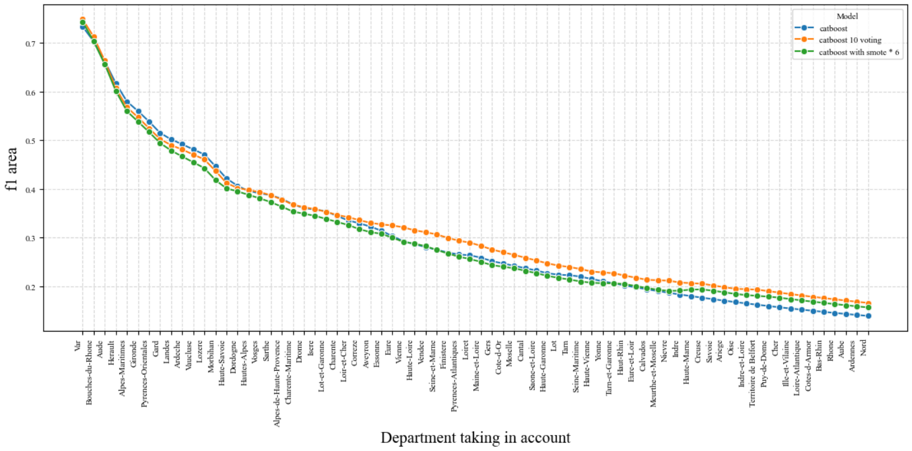

Figure 3: F1 area performance comparison between basic Catboost, with SMOTE (6) and 10 voting models while varying the number of department.

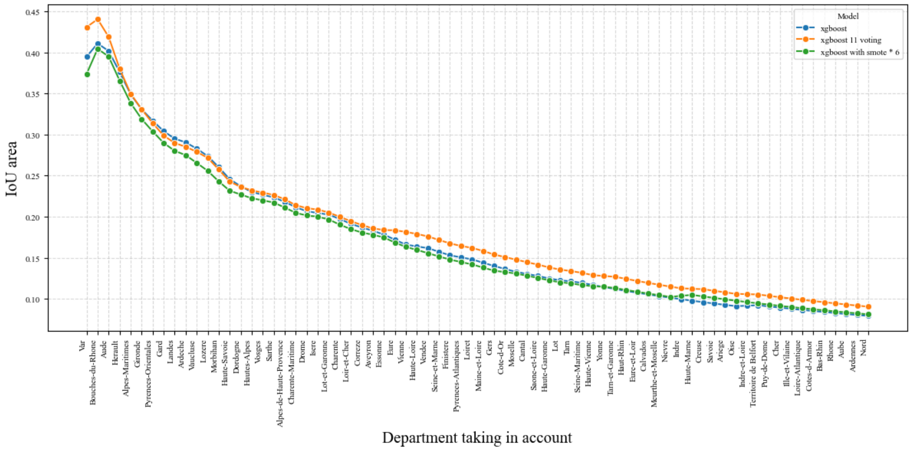

Figure 4: IoU area performance comparison between basic Xgboost, with SMOTE (6) and 11 voting models while varying the number of department.

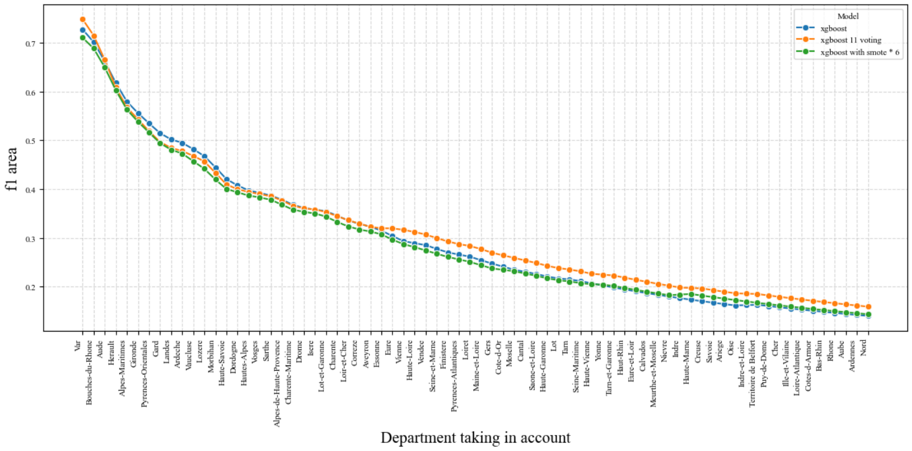

Figure 5: F1 area performance comparison between basic Xgboost, with SMOTE (6) and 11 voting models while varying the number of department.

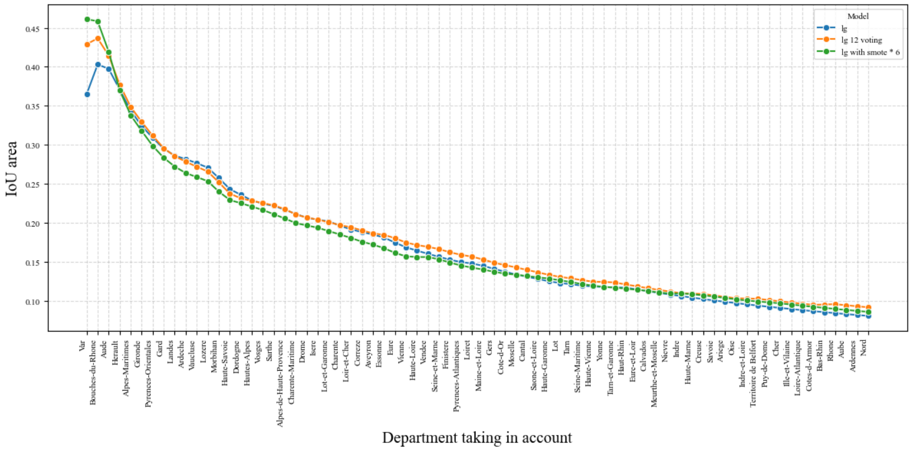

Figure 6: IoU area performance comparison between basic LG, with SMOTE (6) and 12 voting models while varying the number of department.

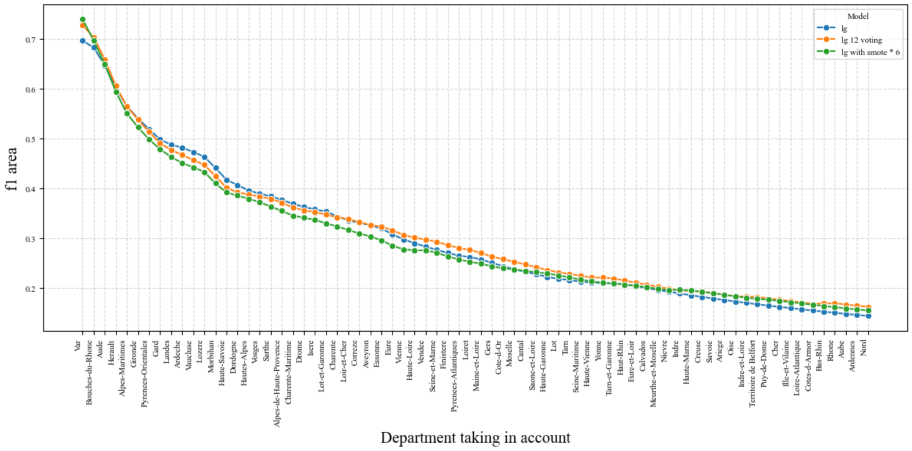

Figure 7: F1 area performance comparison between basic LG, with SMOTE (6) and 12 voting models while varying the number of department.

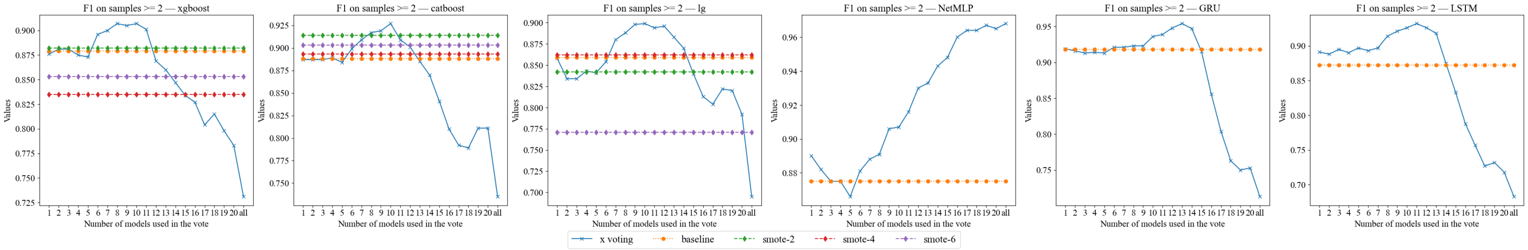

Figure 8: F1 performance on samples of true risk (or predicted) on samples of class superior or equal than 2 between voting, classic and SMOTE models.

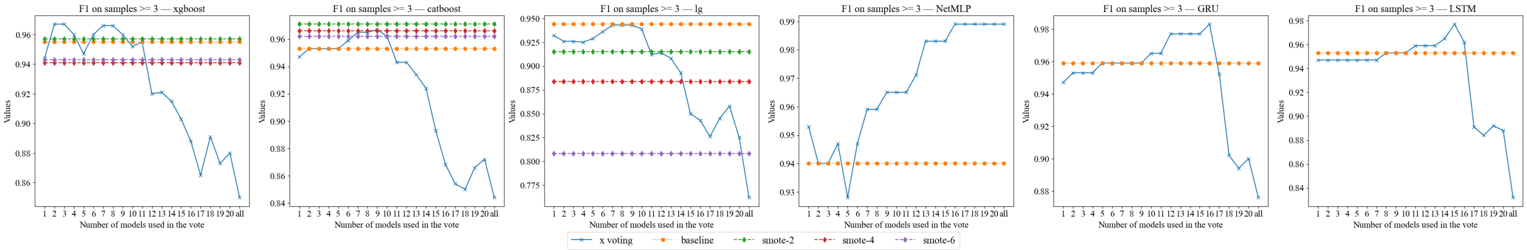

Figure 9: F1 performance on samples of true risk (or predicted) on samples of class 3 or 4 between voting, classic and SMOTE models.