Documentation de ibd2

Nous nous sommes inspirés de la librairie d'explicabilité iBreakDown pour créer un nouvel algorithme traitant lui aussi les modèles non additifs. Nous avons appelé cette librairie ibd2.

La fonction principale de cette librairie est la fonction « compute_explanation_path » qui est appelée par la méthode « explain ».

La fonction compute_explanation_path prend en paramètres en résumé un jeu de données (data), un modèle entraîné à partir de data et un individu (instance) ayant les mêmes variables que data mais n'appartenant pas à data.

Elle renvoie un dictionnaire comportant les variables et leur contribution pour l'individu pris en paramètres (les variables étant classées par contribution décroissante).

Attention, ibd2 ne fonctionne qu'avec une variable à expliquer ayant des modalités binaires (0 et 1).

Explication du fonctionnement de compute_explanation_path :

Exemple avec : data =

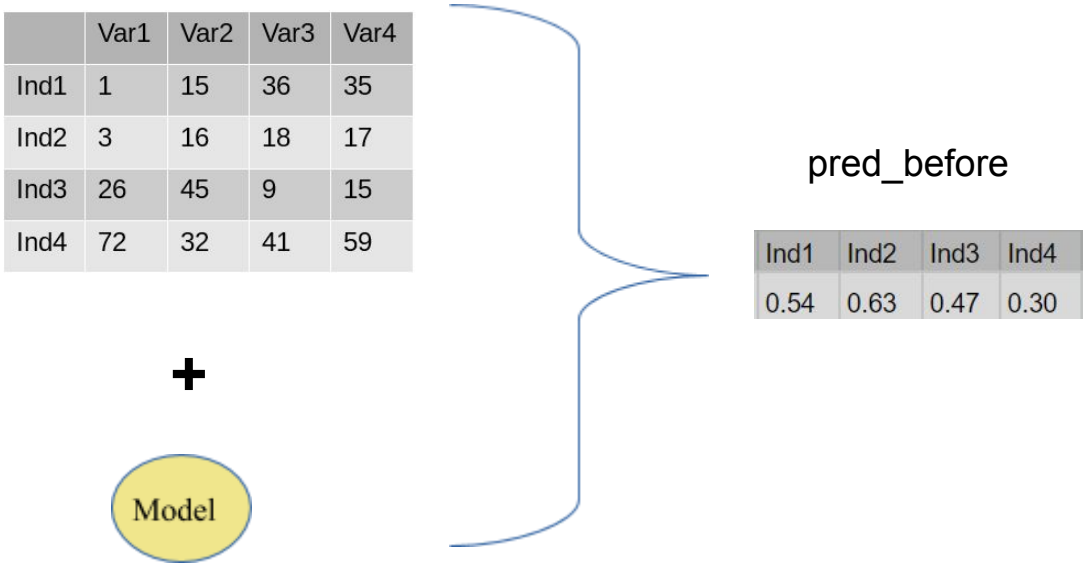
	Var1	Var2	Var3	Var4
Ind1	1	15	36	35
Ind2	3	16	18	17
Ind3	26	45	9	15
Ind4	72	32	41	59

instance =

Var1	Var2	Var3	Var4
15	17	43	36

modèle =

A partir de data et modèle, on commence par calculer la probabilité d'avoir la valeur prédite zéro pour chaque individu de data (pred_before) :



	Var1	Var2	Var3	Var4
Ind1	1	15	36	35
Ind2	3	16	18	17
Ind3	26	45	9	15
Ind4	72	32	41	59

Ind1	Ind2	Ind3	Ind4
0.54	0.63	0.47	0.30

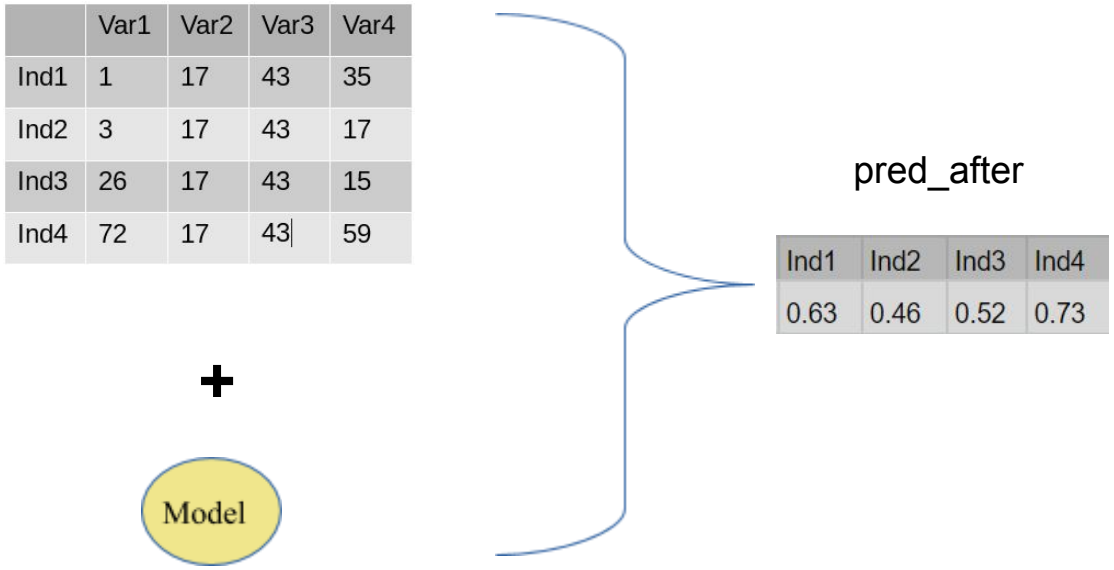
Pour chaque singleton ou doublon de variable ({1}, {2}, {3}, {4}, {1,2}, {1,3}, …), nous faisons une copie de data appelée new_data dans laquelle on remplace les valeurs des variables du groupe par les valeurs de l'instance pour ces mêmes instances.

Exemple : Si groupe = {2,3} :

new_data =

	Var1	Var2	Var3	Var4
Ind1	1	17	43	35
Ind2	3	17	43	17
Ind3	26	17	43	15
Ind4	72	17	43	59

Puis à partir du modèle et de new_data, on prédit à nouveau la probabilité d'avoir la valeur prédite zéro pour chaque individu (pred_after) :



Ensuite, on calcule l'impact en faisant la somme des distances entre les vecteurs pred_before et pred_after.

Si le groupe était un doublon, alors on soustrait à l'impact de ce groupe, l'impact des valeurs seules contenues dans ce doublon.

Puis on range l'impact du groupe dans un dictionnaire.

Fin Pour

Enfin on trie le dictionnaire par ordre décroissant d'impact (=contribution) puis on retourne le dictionnaire.