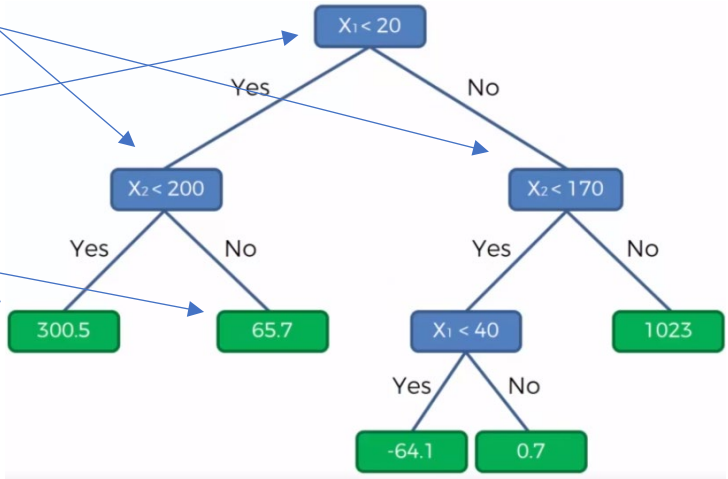
La forêt aléatoire de régression

Le concept et l'algorithme RandomForestRegressor

Le premier concept à expliquer est celui des arbres de régression. L'idée est de d'obtenir la valeur cible en découpant notre jeu de données via des choix binaires.

Ces choix sont appelés des nœuds (nodes) et dépendent de la valeur d'une valeur explicative. Le premier nœud est la racine (root).

L'objectif est de mener vers le résultat prédit selon les valeurs entrées. Ces prédictions s'appellent des feuilles (leaf) et sont le résultat de l'arbre.



Pour définir le choix posé par le nœud, l'algorithme va se placer entre la première et la deuxième valeur de x puis calculer la somme des carrés des écarts entre valeur observée et valeur prédite. La même est faite en plaçant le choix entre la deuxième et troisième valeurs de x et ainsi de suite jusqu'à être entre la n-1ème et énième valeur. Est alors choisi comme choix la valeur pour laquelle la somme des carrés résiduels est la plus petite. Cette méthode est celle des moindres carrés. Une fois le premier nœud posé on répète l'opération pour chaque section jusqu'à ce que chaque feuille soit pure (toutes les valeurs sont correctes) ou contienne un nombre de valeur inférieur au minimum choisi. Il est aussi possible de descendre jusqu'à une profondeur choisie. On se méfie toutefois de conserver suffisamment de valeur dans chaque feuille, généralement il faut une vingtaine de résultats par feuille sinon cela peut signifier que l'algorithme est en « overfitting ».

Maintenant, pour ce qui est de la forêt, l'idée est simplement de répéter la création d'un arbre n fois afin d'affiner le résultat. Dans ce cas, la prédiction est la moyenne des résultats de tous les arbres.

Pourquoi cet algorithme ?

Notre jeu de données contient une trentaine de valeurs explicatives. Cela peut représenter beaucoup pour un algorithme de régression linéaire qui se contentera de prendre certaines de ces valeurs. De son côté, la forêt aléatoire de régression s'adapte à un grand nombre de valeurs explicatives permettant de s'assurer de leur implication ou non dans la valeur cible. De plus, le fait que le résultat soit la moyenne de plusieurs arbres peut limiter les effets de l'overfitting.

Les paramètres, méthodes et attributs de l'algorithme

L'algorithme possède certains paramètres qui nous intéressent :

- n_estimators : il s'agit du nombre d'arbres souhaité dans notre forêt. Sa valeur par défaut est de 100. Dans le cas d'une quinzaine de valeurs explicatives nous avions déjà choisi d'utiliser 100 arbres. Ici nous en utiliserons 150 et pas plus afin d'augmenter la précision en évitant de faire perdre trop de performance à algorithme sans y gagner assez de précision.
- max_depth : il s'agit de la profondeur maximale de nos arbres. Si cette valeur est trop élevée l'algorithme sera surentraîné et incapable de généraliser pour s'adapter à tous les jeux de données.

L'algorithme possède également des attributs dont un nous sera utile à l'analyse du résultat :

- feature_importance_ : il s'agit d'un tableau associant à chaque valeur explicative (feature) une importance sous forme de taux dans le calcul du résultat. La somme des importances donne donc 1.

Enfin, nous retrouvons certaines méthodes que nous utiliserons. Ces méthodes sont les mêmes que pour l'algorithme de régression linéaire veuillez donc vous référer à la veille sur cet algorithme afin d'obtenir plus de détails :

- fit
- score
- predict