3 Encodage

3. Encodage

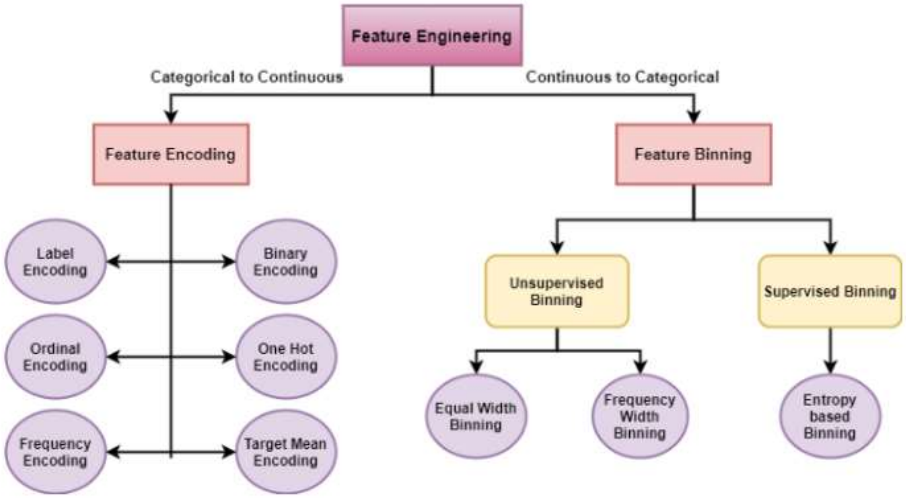
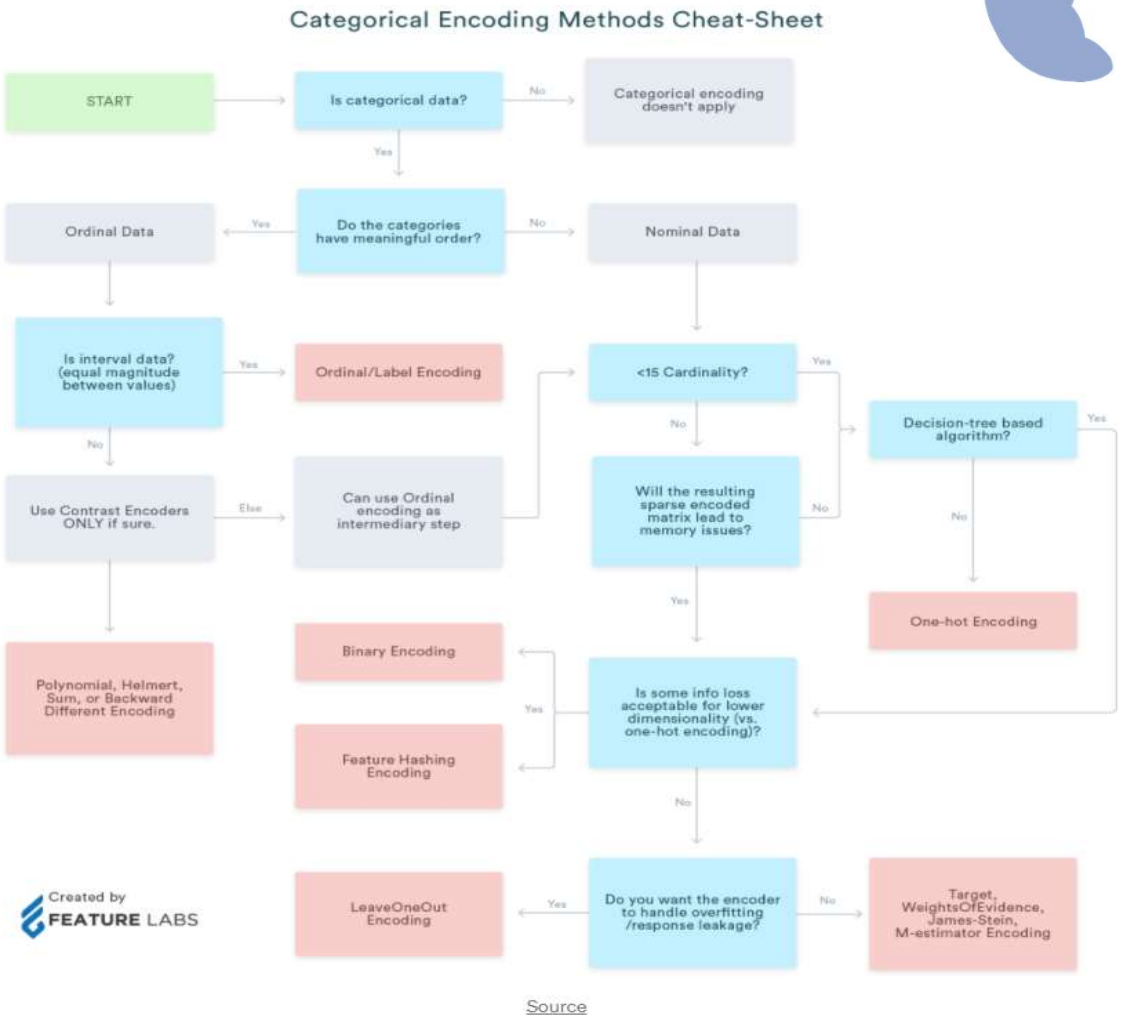


Image by Author

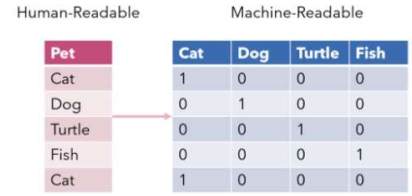


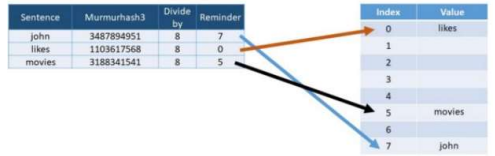
Source

* : sklearn.preprocessing - ** : category_encoders

3. Encodage

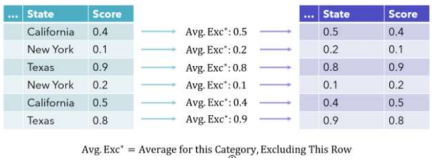
Encodeur	Description	+/-
* OneHotEncoder	connu sous le nom de variables fictives, est une méthode de conversion de variables catégorielles en plusieurs colonnes binaires, où un 1 indique la présence de cette ligne appartenant à cette catégorie.	- : augmente la dimensionnalité mais peu d'information (beaucoup 0), les nouvelles variables en relation linéaire les unes avec les autres ➔ pb parallélisme et multi-colinéarité
** HashingEncoder	Convertir les données en un vecteur de caractéristiques. Cette opération s'effectue à l'aide d'un hachage. Nous appelons cette méthode "hachage de caractéristiques" ou "l'astuce du hachage". Replace var_cat with col_0 to col_nb_vecteur [index 0 to nb_vecteur] which contains either 1 or 0.	+ : si utilisation des modèles d'arbres avec de nombreux niveaux différents. - : interprétabilité difficile de la contribution de chacun de vos niveaux.
* OrdinalEncoder	Nous effectuons un encodage ordinal pour nous assurer que l'encodage des variables conserve la nature ordinale de la variable. Ceci n'est raisonnable que pour les variables ordinales	- : Pas utiliser pour des variables non ordinales
** TargetEncoder	L'encodage de la cible est un moyen très efficace de représenter une colonne catégorique et n'occupe que l'espace d'une seule caractéristique. Également connu sous le nom d'encodage de la moyenne, chaque valeur de la colonne est remplacée par la valeur cible moyenne pour cette catégorie. Cela permet une représentation plus directe de la relation entre la variable catégorielle et la variable cible	- : rend plus difficile pour le modèle l'apprentissage des relations entre une variable codée en moyenne et une autre variable, rend plus difficile pour le modèle l'apprentissage des relations entre une variable codée en moyenne et une autre variable





3. Encodage

Encodeur	Description	+/-
** LeaveOneOutEncoder	Cette méthode est très similaire à l'encodage des cibles, mais exclut la cible de la rangée actuelle lors du calcul de la cible moyenne pour un niveau afin de réduire l'effet des valeurs aberrantes. Comme le modèle est exposé non seulement à la même valeur pour chaque classe encodée, mais aussi à une plage, il apprend à mieux généraliser.	- Puisque chaque valeur de la catégorie est remplacée par la même valeur numérique, le modèle peut avoir tendance à s'adapter de manière excessive aux valeurs encodées qu'il a vue
** CatBoostEncoder	effectue le codage CatBoost pour les caractéristiques catégorielles. Il prend en charge les types de cibles suivants : binaires et continues. Pour le support des cibles polynomiales, il utilise un PolynomialWrapper. Il s'agit d'un codage très similaire à celui du leave-one-out, mais qui calcule les valeurs "à la volée". Par conséquent, les valeurs varient naturellement pendant la phase d'apprentissage et il n'est pas nécessaire d'ajouter du bruit aléatoire.	
** PolynomialEncoder	Le codage polynomial est une forme d'analyse des tendances dans la mesure où il recherche les tendances linéaires, quadratiques et cubiques de la variable catégorielle. Ce type de système de codage ne doit être utilisé qu'avec une variable ordinale dans laquelle les niveaux sont également espacés.	
** JamesSteinEncoder	Comme TargetEncoder. L'encodeur James-Stein évalue la quantité de variation dans les exemples de cette catégorie et la compare à la variation sur l'ensemble des données.	- Distribution normale - S'il n'y a que quelques exemples par catégorie, cette technique ne sera pas particulièrement utile. Nous devons également être conscients que nous devons "sauvegarder" la moyenne du groupe pour chaque catégorie



* : sklearn.preprocessing - ** : category_encoders

3. Encodage

Encodeur	Description	+/-
** HelmertEncoder	Pour la valeur de la caractéristique, l'estimateur de James-Stein renvoie une moyenne pondérée de : La valeur cible moyenne pour la valeur de la caractéristique observée. La valeur cible moyenne (indépendamment de la valeur de la caractéristique). L'encodeur James-Stein rétrécit la moyenne vers la moyenne générale. Il s'agit d'un codeur basé sur la cible.	

* : sklearn.preprocessing - ** : category_encoders