Notre équipe 19 lors du Hi Hackathon

TRAN-THUONG Tien-Thinh

Abstract

Dans notre équipe 19, nous sommes parti sur différents algorithmes. Il s'est avéré que XGBoost a été la meilleure solution. Pour ma part j'ai travaillé sur les réseaux de neurones avec les modules Tensorflow et Keras.

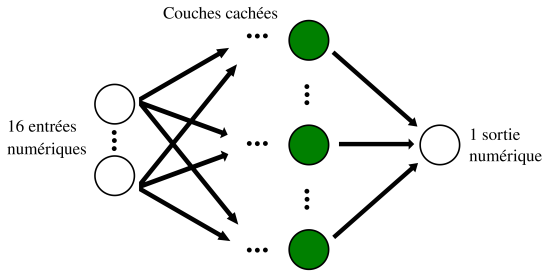
Description de ce que j'ai fait

J'ai tout d'abord fait du pré-processing sur les données, avec l'ensemble de l'équipe, nous avons discuté des variables les plus utiles et comment nous allions supprimer les outliers. J'ai pu apprendre à utiliser le $z - score$.
Puis j'ai appliqué mon réseau de neurones directement sur mes données numériques, avec un résultat d'$explained\_variance\_score$ de 51%.

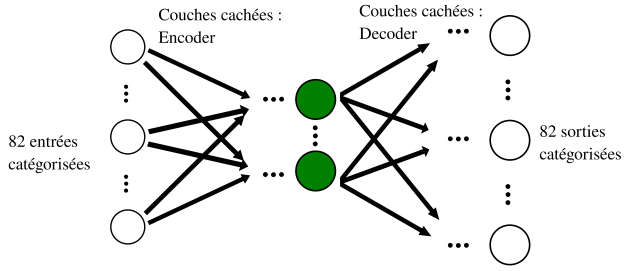
Ce que j'ai fait pour aller plus loin

J'ai réalisé un modèle que je n'ai pas pû entrainer jusqu'au bout par manque de temps. Permettez-moi de vous le présenter tout de même.

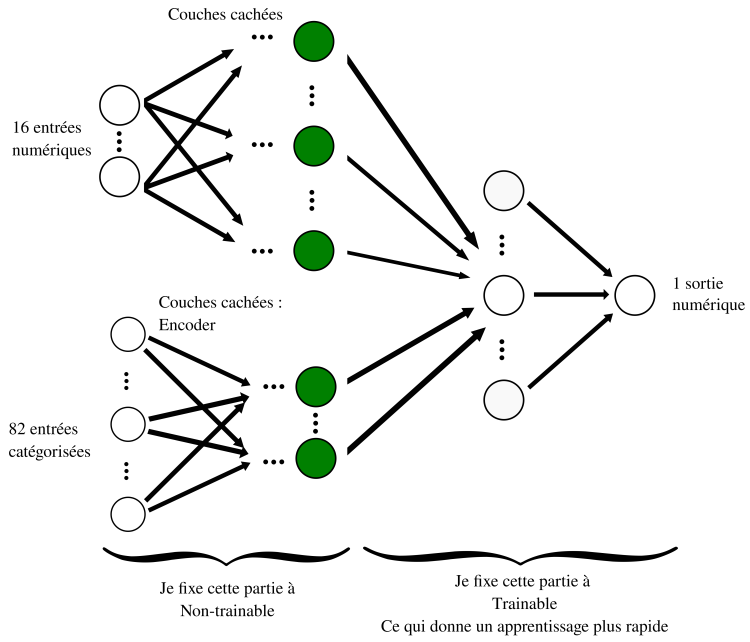
1 - J'ai voulu créer un modèle simple qui traite les valeures numériques et qui prédit la consommation annuelle (activation: relu)



2 - J'ai voulu créer un Encoder-Decoder afin de pré-entrainer le modèle à comprendre les variables catégorisées (activation: sigmoïde)



3 - J'ai voulu concaténer les 2 modèles précédents (fixés à non-trainable) afin de tirer parti de leur compréhension des données. C'est selon moi une sorte de Transfert d'apprentissage.



La difficulté que j'ai repérée dans la base de données était le mélange entre données *numérique* et *catégorisées* ainsi que la quantité importante des données. J'ai donc entrainé deux modèles séparément, le premier sur les *données numériques* sur le modèle d'une régression linéaire, et le second sur les *données catégorisées* suivant le modèle d'*Encoder-Decoder*.
Une fois les deux modèles entrainés, je retire les couches après la couche de neurones *vertes* sur le schéma. Je suppose alors que les couches de neurones devant la couche de neurones *vertes* ont une bonne compréhension des données en entrées. Je n'ai alors qu'à utiliser leurs sorties telles quelles et entrainer les deux dernières couches de neurones à prédire la consommation annuelle.

En procédant ainsi, j'ai pu réduire le nombre de paramètres à entrainer à chaque fois et j'ai également séparé le traitement des données *numérique* et *catégorisées*.