Projet de fin d'étude

Rétro-ingénierie d'un modèle déterministe

Aymeric Duchein - Benoit Pimpaud - Jordan Rostren - Paul Schaeffer

Sommaire

- Introduction et contexte
- Analyse des données
- Data processing
- Modélisations
- Critiques et améliorations
- Conclusion

Introduction et contexte

Contexte et rétro-ingénierie

- Estimer les frais d'acquisitions lors de l'achat/location d'un bien immobilier
- Données issue d'un modèle déterministe
- Rétro-ingénierie du modèle avec du machine-learning

NEXVIA
real estate at heart

Environnement technique

Traitements et analyses

- Numpy, Pandas
- Searborn

Machine Learning

- Scikit-learn et Keras

Intégration continue

- Circle CI
- AWS S3

Gestion de projet et “industrialisation”

- Gestion de version et développement en groupe avec **Git**
- Tests unitaires et intégration continue avec **Circle CI**
- **Règles stylistiques** en respectant la norme PEP8

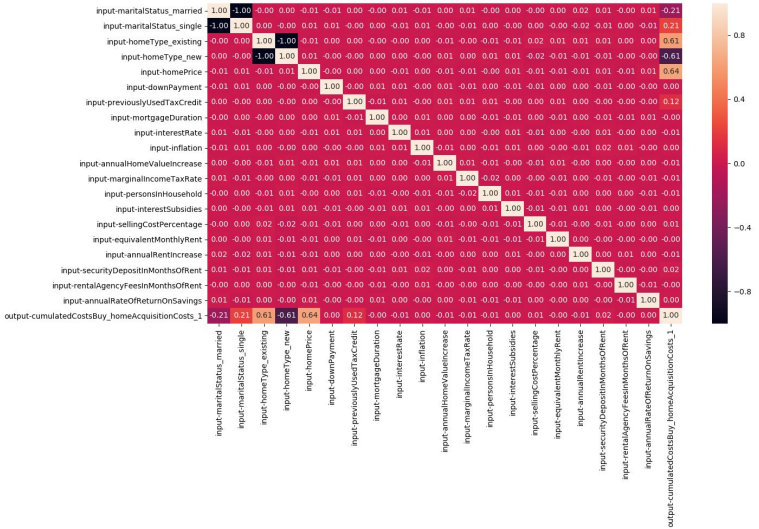
Analyses des données

Les variables

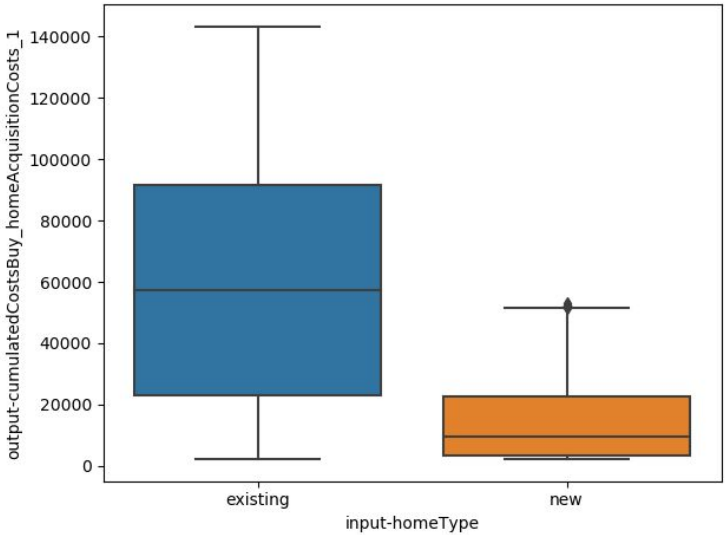
Une vingtaine de variables dont :

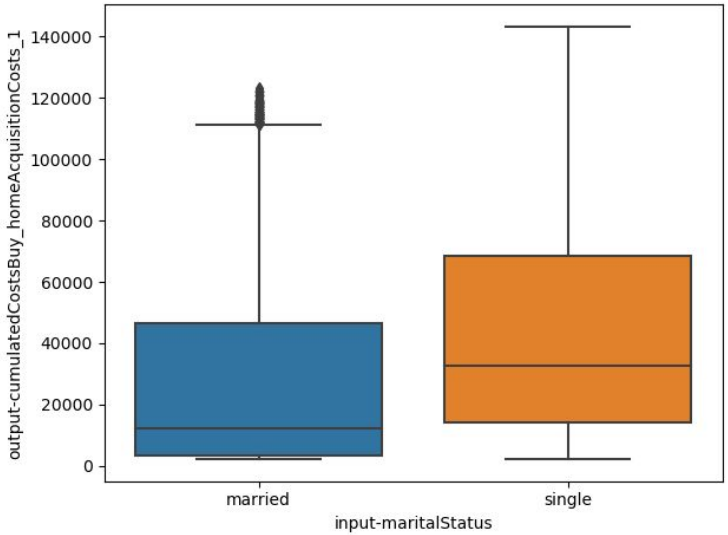
- Le prix du bien
- Le type de bien (nouveau/ancien)
- La situation marital
- Les crédits d'impôts
- Le taux d'imposition
- etc...

Corrélations



Distributions des variables binaires





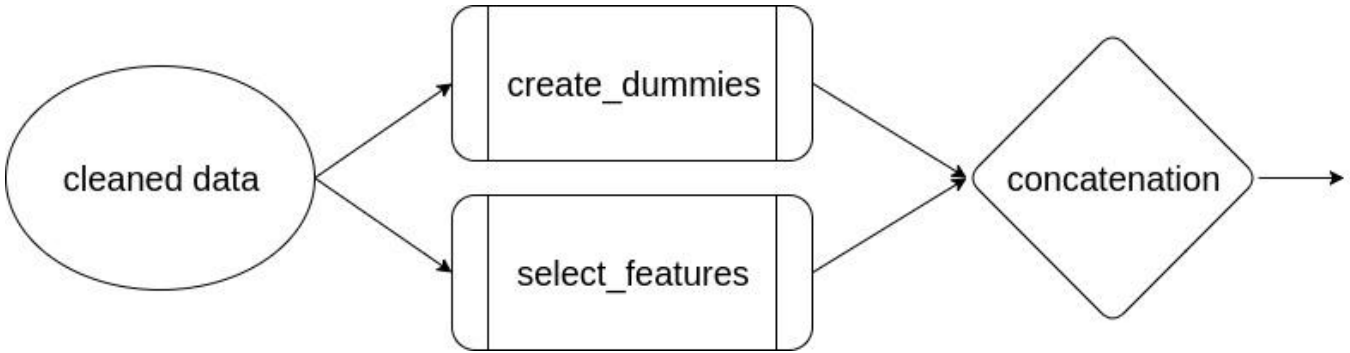
Data processing

Encodage des variables qualitatives

input-marital-status
single
single
married

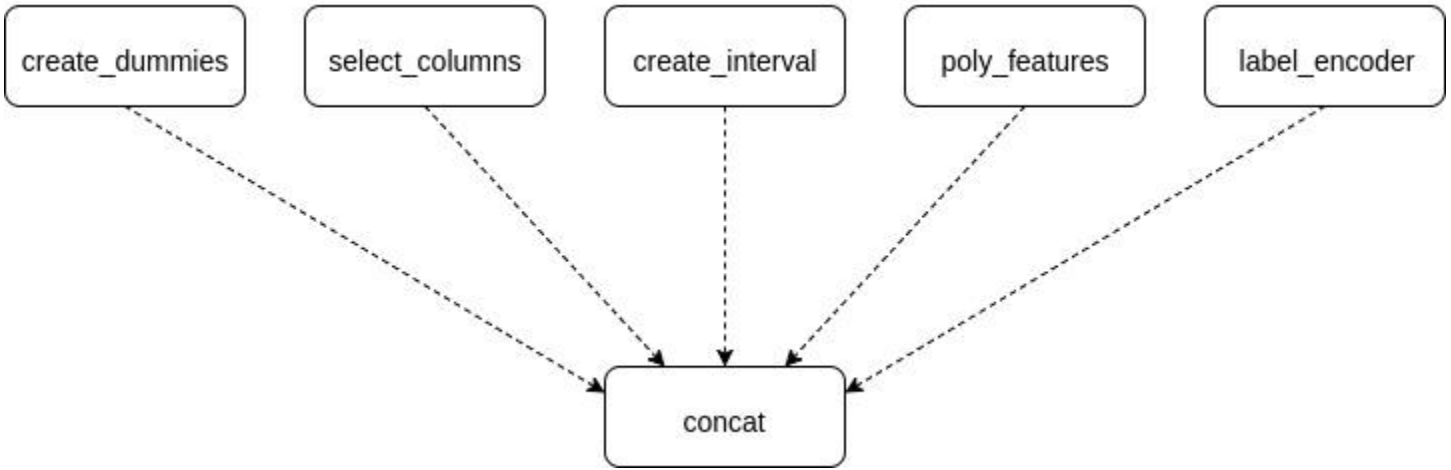
single	married
1	0
1	0
0	1

Exemple d'une fonction de processing



processing_full

Processing modulaire



Modélisations

Evaluation des modèles

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i|$$

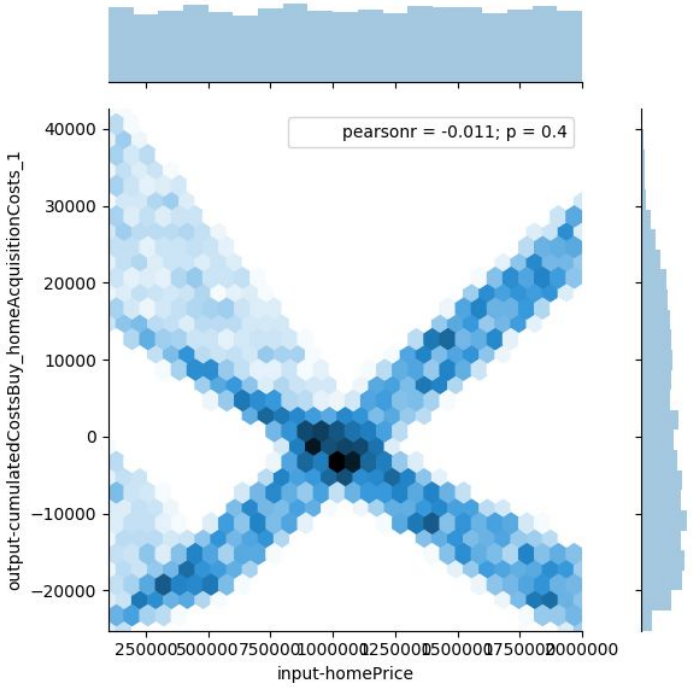
$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$$

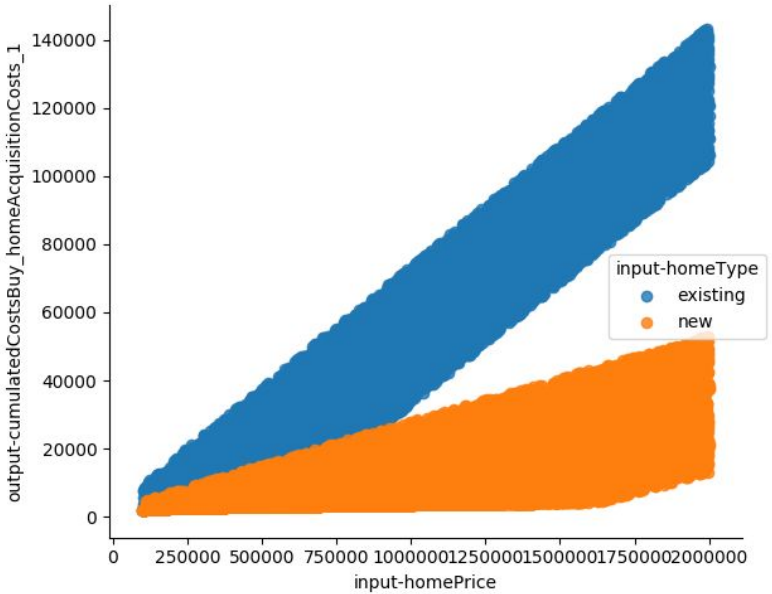
Régression linéaire

$$y = \sum a_i . x_i$$

MAE	RMSE	R2
12 000	15 000	0,833



Régressions linéaires selon le type de bien



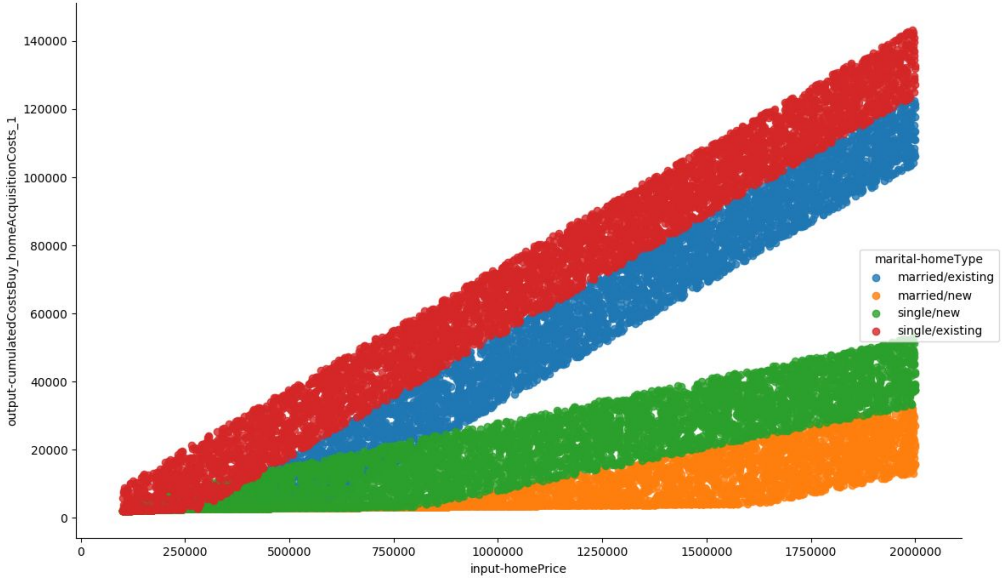
Régressions linéaires selon le type de bien

$$_{new}(x) = \sum \alpha_i x_i \ , \ m_{existing}(x) = \sum \beta_i x_i$$

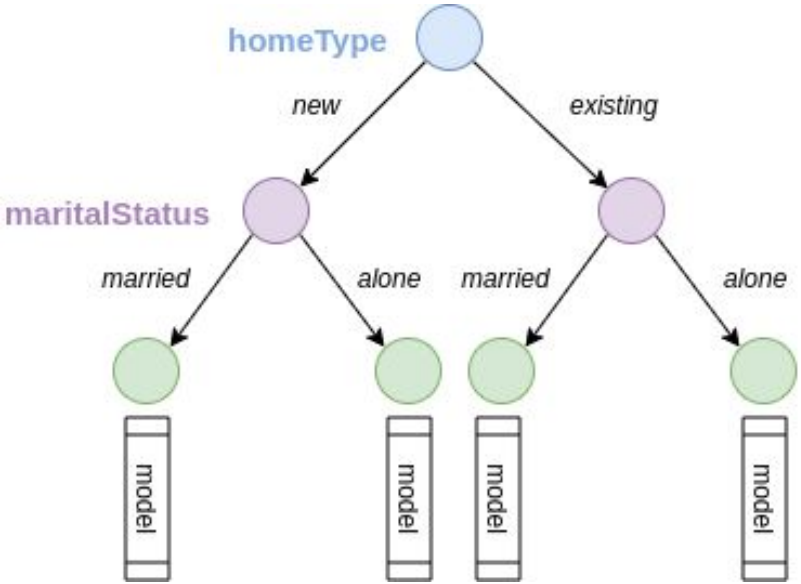
$$\mathrm{f(x)} = \mathrm{m}_{new}(x).x_{|homeType=new} + m_{existing}(x).x_{|homeType=existing}$$

MAE	RMSE	R2
3 100	4 200	0,986

Split Discret Features Model (SDFM)



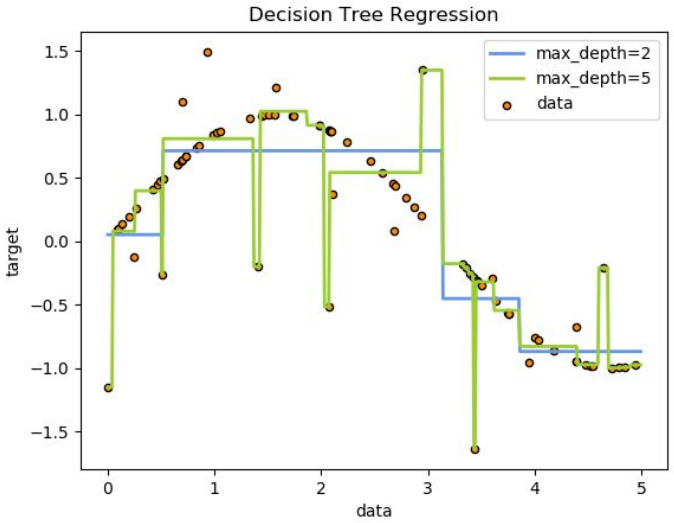
Split Discret Features Model (SDFM)

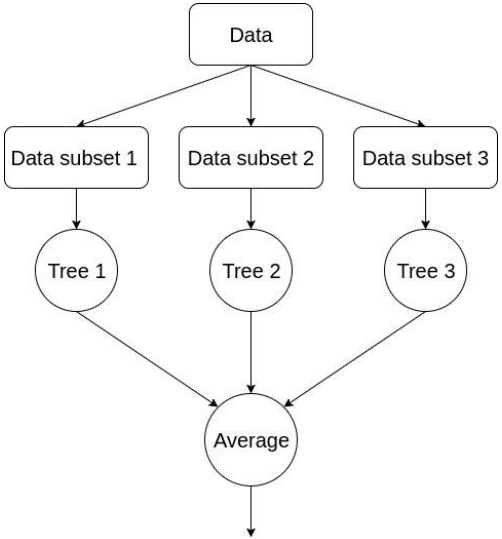


MAE	RMSE	R2
2 000	3 100	0,9927

Pour des modèles de régressions

Arbres de décisions et Random Forest





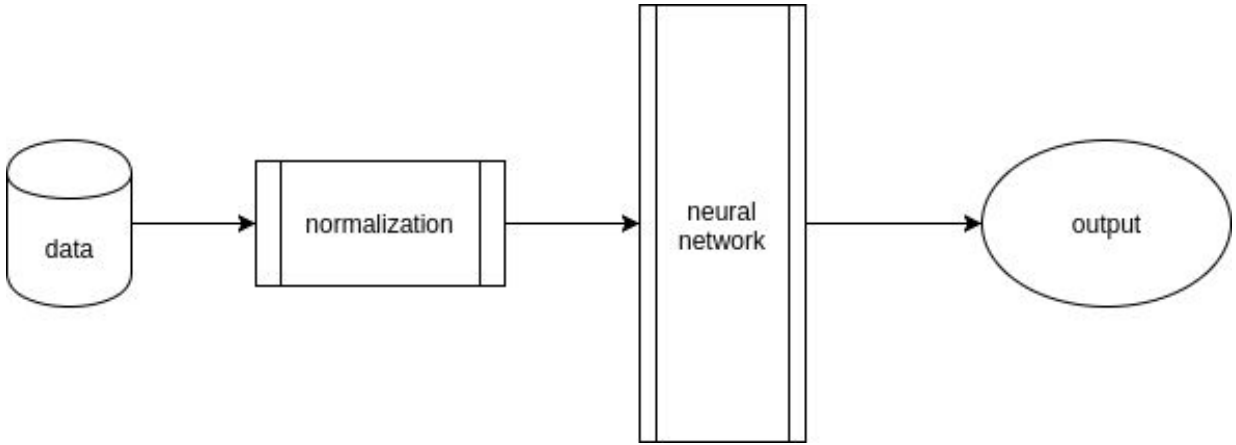
Arbres de décisions et Random Forest

Modèle	MAE	RMSE	R2
Arbre de décision	700	1000	0,99921
Random Forest	300	450	0,99984

Réseaux de neurones

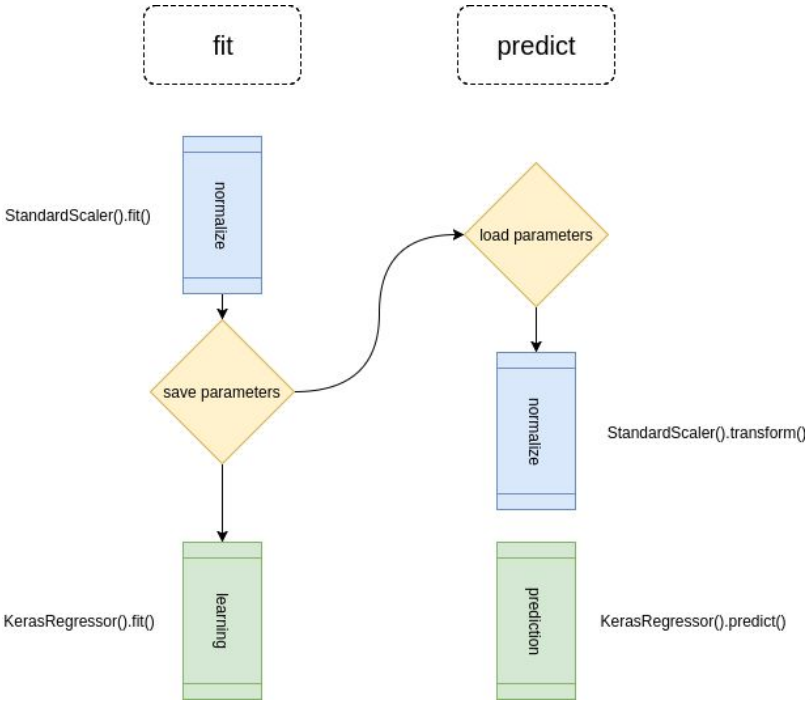
- Processing & Workflow
- Architectures
- Résultats

Normalisation

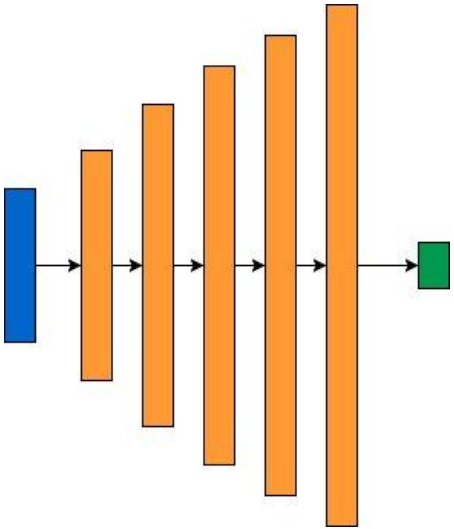


Encapsulation avec Keras

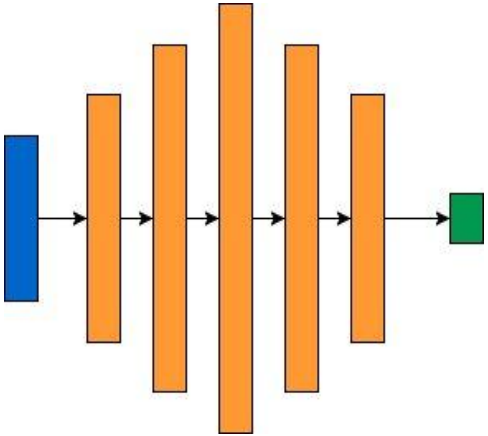
- Interface Keras - scikit-learn
- Utilisation des Pipelines scikit-learn



Architectures

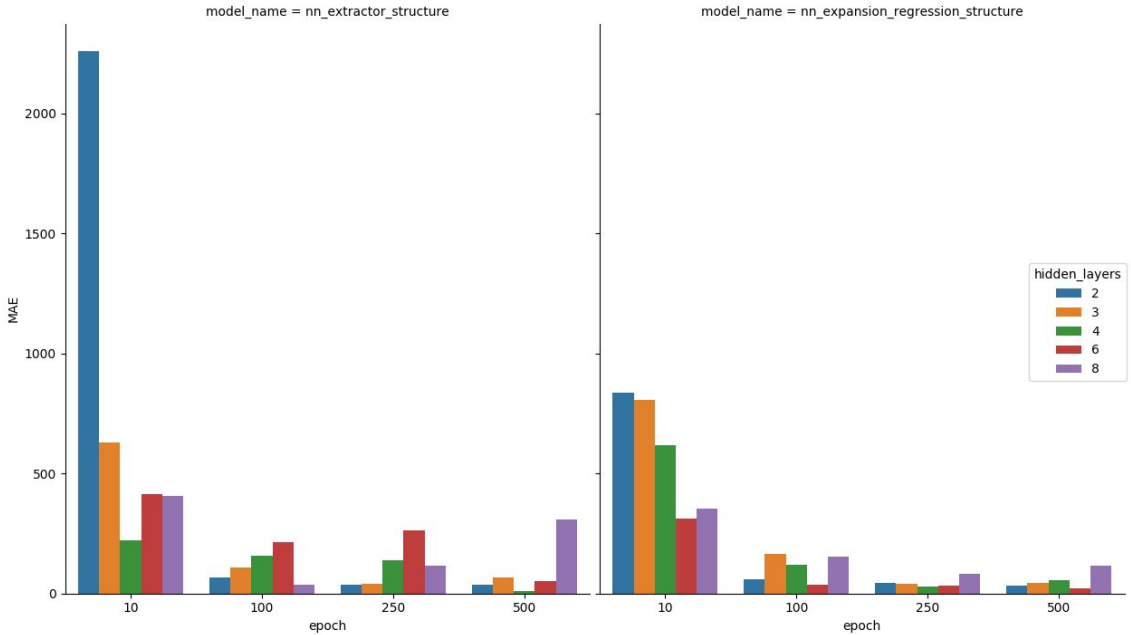


extraction



expansion-regression

Résultats


model_name = nn_extractor_structure
model_name = nn_expansion_regression_structure
MAE
2000
1500
1000
500
0
10
100
250
500
epoch
hidden_layers
2
3
4
6
8

Critiques et améliorations

Modèle	MAE	RMSE	R2
Régression linéaire	12 000	15 000	0,833
Régression selon homeType	3 100	4 200	0,986
SDFM (régression)	2 000	3 100	0,9927
Arbre de décision	700	1000	0,99921
Random Forest	300	450	0,99984
Réseaux de neurones (entrainement court)	30	50	0,99999
Réseaux de neurones (entrainement long)	6	8	0,99999

Conclusion

Merci de votre attention