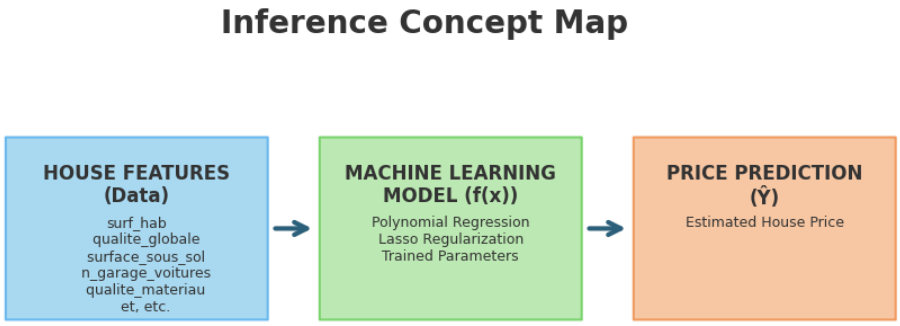
Prédiction des Prix Immobiliers : Une Approche Comparative

Analyse et Sélection de Modèles de Régression

Table of contents

1 Présentation du Problème

Notre objectif est de prédire le prix de vente de maisons (problème de **régression supervisée**) à partir de leurs caractéristiques, en utilisant un modèle capable d'inférer une valeur continue.



---

2 Données et Modèles Testés

Les données utilisées proviennent d'un jeu de données immobilier, comprenant des caractéristiques telles que la surface, le nombre de pièces, l'année de construction, l'emplacement, etc. Chaque ligne correspond à une maison, avec son prix de vente comme variable cible.

Nous avons évalué plusieurs modèles de régression, en augmentant progressivement leur complexité :

- **Baseline :** Régression Linéaire simple (`LinearRegression`), qui sert de point de comparaison.
- **Régression Polynomiale :** Ajout de termes polynomiaux (`PolynomialFeatures`) pour des degrés de 2 à 6, afin de capturer des relations non linéaires.
- **Modèles Régularisés :** Combinaison de la régression polynomiale avec des techniques de régularisation :
- **Ridge (L2)** et **Lasso (L1)**, pour des degrés de 2 à 6 et des valeurs de `alpha` de 0 à 1000.

---

3 Analyse des Résultats

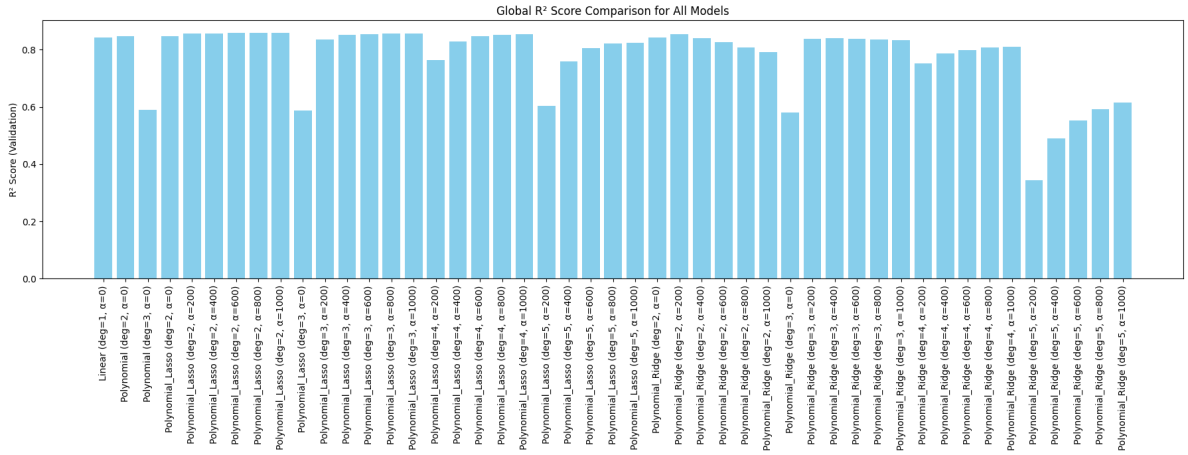
Les graphiques ci-dessous présentent les performances des modèles selon deux métriques : - **R^2 (score de détermination)** : mesure la qualité de la prédiction (plus proche de 1 = meilleur). - **MAE (erreur absolue moyenne)** : mesure l'écart moyen entre les prix réels et prédits (plus bas = meilleur).

Top 5 configurations R^2 : - Les meilleurs scores R^2 sont obtenus avec le modèle **Polynomial Lasso (degré=2, alpha=1000 à 200)**, atteignant jusqu'à **0.859**. - À titre de comparaison, le modèle ayant la plus faible MAE (**Polynomial Ridge (degré=2, alpha=200)**) affiche un R^2 de **0.854**, soit très proche du meilleur modèle R^2.

Top 5 configurations MAE : - Le modèle **Polynomial Ridge (degré=2, alpha=200)** obtient la plus faible MAE (**$18,410**), ce qui signifie qu'il minimise le coût moyen des erreurs de prédiction. - À titre de comparaison, le meilleur modèle R^2 (**Polynomial Lasso (degré=2, alpha=1000)**) a une MAE de **$18,717**, soit une différence faible par rapport au meilleur MAE.

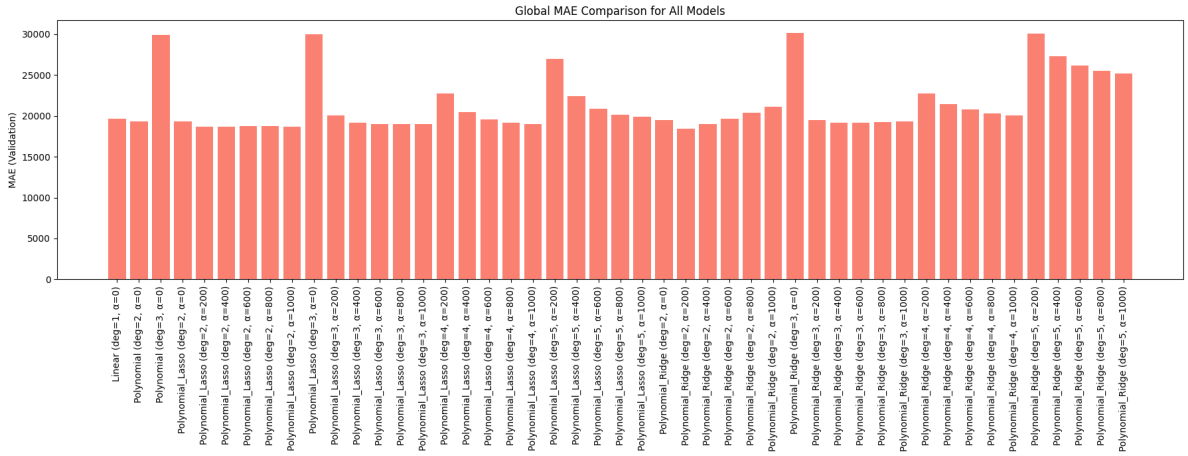
Choix du modèle final : Les deux modèles sont très proches en termes de R^2 et de MAE. J'ai choisi le modèle avec la plus faible MAE (**Polynomial Ridge (degré=2, alpha=200)**), car la valeur R^2 des deux modèles était quasiment identique, et minimiser l'erreur absolue moyenne est plus pertinent pour limiter l'impact financier des erreurs de prédiction.

3.1 Graphs de comparaison des modèles



Ce graphique compare les scores R^2 des différents modèles testés, illustrant leur capacité à expliquer la variance des prix immobiliers.

3.2 Graphs de comparaison des modèles



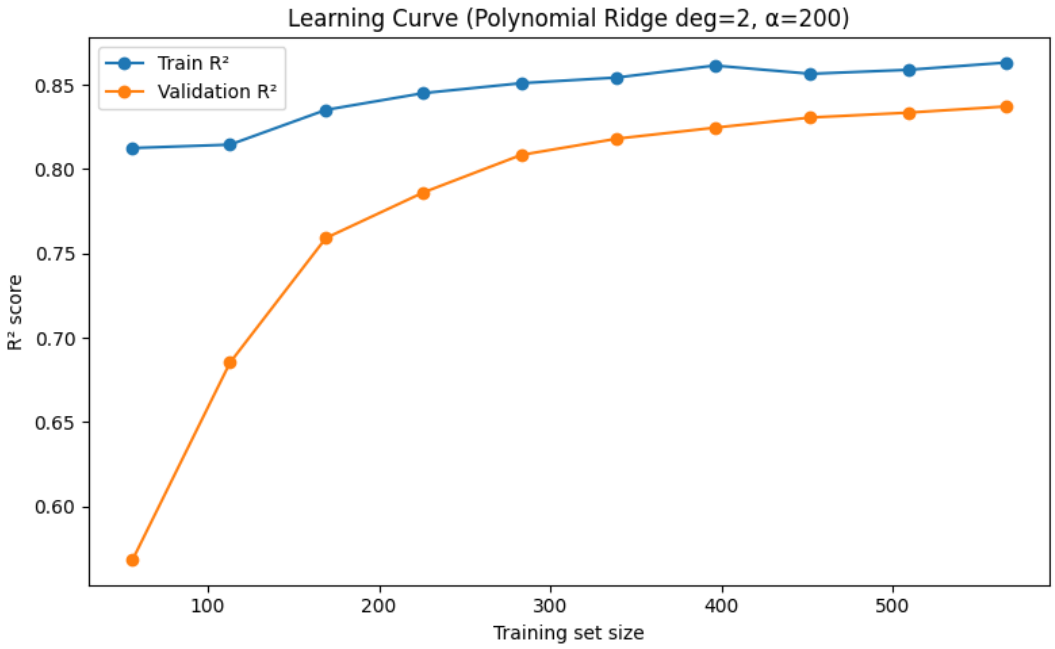
Ce graphique présente la comparaison des erreurs absolues moyennes (MAE) pour chaque modèle, permettant d'évaluer la précision des prédictions.

---

4 Choix de l'Ordre du Polynôme et Learning Curve

L'analyse comparative montre que le **degré polynomial 2** offre le meilleur compromis performance/complexité, les degrés supérieurs n'apportant pas d'amélioration significative.

*La **Learning Curve** ci-dessous confirme que notre modèle de degré 2 généralise bien : les scores d'entraînement et de validation convergent vers une performance élevée, indiquant une faible variance (pas de surapprentissage).*



---

5 Présentation de la Régularisation Ajoutée

Pour éviter le surapprentissage, nous avons comparé deux techniques de régularisation : Ridge et Lasso.

- **Lasso (L1) :** Pénalise la somme des valeurs absolues des coefficients . Peut réduire des coefficients à zéro, réalisant ainsi une **sélection de variables**.
- **Ridge (L2) :** Pénalise la somme des carrés des coefficients. Réduit la magnitude des coefficients pour gérer la multicolinéarité.

Nous avons choisi **Ridge**, car il a produit une **erreur absolue moyenne (MAE) plus faible**, ce qui est plus pertinent pour évaluer l'impact monétaire des erreurs de prédiction.

---

6 Meilleur Modèle : Choix et Résultats Finaux

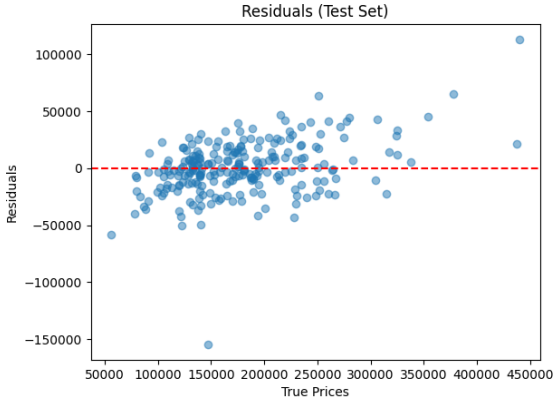
Le meilleur modèle sélectionné est `PolynomialFeatures(degree=2) + Ridge(alpha=200)`.

- **Fonction Coût (Ridge) :** Ce modèle minimise la somme des carrés des résidus (RSS) plus une pénalité L2
- **Comparaison Finale (Validation Set) :**

Modèle	R^2	MAE
Polynomial Ridge (deg=2, =200)	**0.854**	**$18410**
Polynomial Lasso (deg=2, =1000)	0.859	$18717
Régression Linéaire Simple	0.84	$19645

6.1 Résultats sur le Jeu de Test

Le modèle final montre d'excellentes performances de généralisation.



- mae: 18410
- r2_score: 0.854