Institut Supérieur des Sciences Appliquées et Économiques

associé au Conservatoire National des Arts et Métiers

Analyse Multivariée Approfondite

Projet STA 201

Nom du projet :

« Prédictions des prix immobiliers »

Prépare par :

« Elio Bou Serhal »

Professeurs responsables :

« Dr Zainab Assaghir »

« Dr Amal Kobeissi »

Date : 24-7-2024

Contents

I. Préface et Remerciements

Ce projet est le fruit de plusieurs mois de travail et de recherches intensives. Premièrement, Je tiens à exprimer ma profonde gratitude et ma reconnaissance à mon professeur « Zeinab Alsaghir » pour ces conseils, ces efforts d'explications et son soutient tout au long de ce projet.

Deuxièmement, Je remercie également les membres de l'université CNAM- Liban pour leurs enseignements et leurs soutiens académiques. Enfin, Je remercie mes collègues, amis et parents pour leur discutions, motivations et les encouragements tout au long de cette période.

II. Résumé

En finance et en Economie, la prédiction des prix immobiliers est un domaine crucial qui affecte soit positivement ou négativement sur les décisions d'achat, de vente et d'investissement mais, les progrès d'apprentissage automatique nous permettent maintenant d'utiliser des différents modèles pour faire la prédiction des tendances du marché immobilier.

L'urbanisation et l'augmentation de la population ont conduit à une demande de logement, impactant directement les prix, c'est pour cela vous pouvez voir les acteurs du marché, qu'ils soient acheteurs, vendeurs ou investisseurs prennent des décisions d'investissements basées sur les méthodes comparatives en analysant les données historiques et en utilisant des modèles de régressions linéaires.

III. Liste des Figures

IV. Listes des Tableaux

V. Glossaire

Tableau 1: « Tableau contenant les définitions de chaque mot technique utilisé »

Termes	Définitions
« Régression Linéaire »	
Régression Linéaire	Méthode statistique pour modéliser la relation entre une variable cible et une/plusieurs variables indépendantes (explicatives).
Régression Linéaire (Ascendante)	Méthode de régression ajoutant chaque fois des variables explicatives pour optimiser le modèle.
Régression Linéaire (Descendante)	Méthode de régression éliminant chaque fois des variables explicatives pour optimiser le modèle.
OLS	Régression des Moindre carres Ordinaire qui minimise la somme des carres des différences entre les valeurs observées et valeurs prédites
PCR	Méthode de régression qui utilise les composantes principales pour réduire la dimensionnalité des variables explicatives.
PLS	Méthode de régression qui optimise la covariance entre les variables explicatives et la variable cible.

LASSO	Méthode de régression qui ajoute la somme des valeurs absolues des coefficients à la fonction du prix pour favoriser la sparsité et sélectionner les variables importantes.
Ridge	Méthode de régression qui ajoute la somme des carrés des coefficients à la fonction du prix, réduisant la variance sans éliminer des variables pour régulariser le modèle et réduire la colinéarité.
Elastic Net	Méthode de régression combinant les méthodes de régularisation (LASSO & Ridge) pour équilibrer la sparsité et la régularisation.
Sparsité	Une petite partie des variables explicatives ont des coefficients non nuls qui contribuent aux prédictions des modèles.
« Mesures Métriques »	
MSE, MAE, RMSE, R^2, R^2 adj, AUC	Mesures des erreurs calculées, utilisées pour évaluer la performance des modèles de régression.
CIA, CIB	Critères utilisés pour comparer la qualité des modèles statistiques ayant trop de paramètres.
« En General »	
Imputation	Technique utilisée pour remplacer les valeurs manquantes NA.
CV	Technique d'évaluation d'un modèle en le testant sur des sous-ensembles de données non utilisés lors de la prédiction.
VIF	Mesure de la colinéarité des variables explicatives.
Homoscédasticité	Lorsque la variance des erreurs de prédiction est constante à travers toutes les valeurs de la variable explicative.
Hétéroscédasticité	Lorsque la variance des erreurs de prédiction varie à travers les valeurs de la variable explicative.
Multicolinéarité	Lorsque deux ou plusieurs variables explicatives dans un modèle sont fortement liées, rendant difficile l'estimation des coefficients.
Data_transformed_one_hot	Enjeux de données transformer en utilisant le codage One-Hot
Data_transformed_label	Enjeux de données transforme en utilisant le codage d'Etiquette

VI. Listes des Abréviations

Tableau 2 : « Tableau contenant les définitions sur chaque abréviations »

Termes	Définitions
« Modèles de Régressions »	
RL	Régression Linéaire
PCR	Régression sur les composantes principales (Principale Component Regression)
PLS	Régression des Moindres carrés partiels (Partial Least Squares Regression)
OLS	Régression des moindres carres ordinaires (Ordinary Least square Regression)
« Modèles de Régularisations »	

Ridge	Régression de Ridge (Ridge Regression)
LASSO	Operateur de Régression et de Sélection Absolue Minimale (Least Absolute Shrinkage & Selection Operator)
NE	Relastique net (Relastic Net)
« Mesures Métriques »	
MSE	Erreur Quadratique Moyenne (Mean Square Error)
MAE	Erreur Absolue Moyenne (Mean Absolute Error)
RMSE	Racine de l'Erreur Quadratique Moyenne (Root Mean square Error)
R^2	Coefficient de Détermination (Coefficient of Determination)
$R^2\ adj$	Coefficient de Détermination Ajusté (Adjusted Coefficient of Determination)
AUC	Aire sous la Courbe (Area Under Curve)
CIA	Critère d'information d'Akaike (Akaike Information Criterion)
CIB	Critère d'information Bayésien (Bayesian information criterion)
« En General »	
IA	Intelligence Artificielle (Artificial Intelligence)
AM	Apprentissage Automatique (Machine Learning)
FIV	Facteur d'Inflation de la Variance (Variance Inflation factor)
CV	Validation Croisée (Cross validation)

VII. Introduction

1. Introduire le Problème

1.1) Qu'est l'objectif de cette étude ?

L'objectif de cette étude c'est d'utiliser des modèles de prédictions pour estimer les prix des propriétés immobiliers à partir d'une base de données détaillées sur les maisons dans la ville américaine Ames, Lowa. En utilisant les différents modèles de régression linéaires et de régularisation nous allons identifier les caractéristiques les plus importants pour le prix de ventes des maisons. Donc, cette étude vise à déterminer les variables les plus significatives et d'améliorer la prédiction dans le secteur immobilier.

1.2) Quel est la problématique ?

La problématique de cette étude c'est de prédire le prix des maisons avec des différents conditions : Le grand nombre de variables explicatives, Le type des variables explicatives (quantitatives et qualitatives) ainsi que, les problèmes de la multi colinéarité et de surajustement des modèles. De plus, l'imputation des valeurs manquantes et le choix des meilleures méthodes de régression pour optimiser les prévisions.

1.3) Quelle est la méthodologie de cette étude ?

La méthodologie de cette étude suit les étapes suivantes :

- Etape 1 : **Importation et préparation des données** : Nous allons charger la base de données et appelé les librairies nécessaires sur Rstudio pour commencer à travailler.

- Etape 2 : **Analyse descriptive** : Nous allons explorer les données (structure, dimensions, existence des valeurs manquants et des duplications).
- Etape 3 : **Prétraitement des données** : Nous allons nettoyer la base de données des valeurs manquantes et des duplications. Apres, nous allons transformer les variables qualitatives en variable quantitatives ensuite, nous allons vérifier les conditions des modèles de régression avant de commencer aux appliques.
- Etape 4 : **Partitionnement des données** : Nous allons diviser la base de données en deux ensembles aléatoirement, l'ensemble d'entrainement qui représente (70% - 80%) de la population et l'ensemble de test qui représente (20% - 30%) de la population.
- Etape 5 : **Explication Théorique** : Nous allons expliquer simplement l'algorithme de chaque modèles appliques. En outre, nous allons définir les mesures métriques pour étudier la performance des modèles.
- Etape 6 : **Application des Modèles** : nous allons entrainer et évaluer les divers modèles de régression avances y compris la régression linéaire (RL), la régression des composantes principales (PCR), la régression par moindres carres partiels (PLS) et les modèles de régularisations (Ridge, LASSSO et Elastique Net).
- Etape 7 : **Analyse des Résultats** : Nous allons comparer les performances des modèles en utilisant des mesures métriques telles que le MSE, MAE, RMSE, R^2, R^2 adj et AUC.

2. Décrire et justifier la méthode

2.1) Quelles sont les méthodes utilisées dans notre étude ?

Dans cette étude, nous allons utilises plusieurs techniques de régression y compris la régression linéaire avec l'approche Ascendante **(Forward sélection)** et la régression linéaire par moindres carres ordinaires **(OLS)**, ainsi que la régression par composantes principales **(PCR)** et par moindre carrées partiels **(PLS).** De plus on va appliquer les modèles de régularisation come **Ridge, LASSO**, et **Elastic Net**.

2.2) Pourquoi nous avons sélectionné ces méthodes ?

Nous avons sélectionné ces modèles à cause de leur efficacité a géré des jeux de données complexes avec un grand nombre de variables mixés. Autrement dit ces modèles nous permettent d'identifier des relations entre les différentes caractéristiques des propriétés et leurs prix alors, ces modèles résolvent les problèmes de multi colinéarité et évitent le surapprentissage

3. Annoncer le Plan

Premièrement nous présenterons les enjeux du problème, **Deuxièmement** nous proposerons une revue de la littérature pertinente sur le sujet et **Troisièmement**, on décrira les méthodes utilises dans cette étude puis, les résultats seront présents et discutes avec des comparaisons en **Quatrième** partie. Pour finir, la **Cinquième** partie contiendra les conclusions principales et identifiera à la fois les limites de l'étude ainsi que des pistes de recherche complémentaire.

VIII. Développement – corps du mémoire

1. Matériels et Méthodes

1.1) Quel est le matériel de l'étude ?

Notre base de données contient des informations détaillées sur 2921 maisons dans la ville d'Ames, Lowa. Cette base de données est formée de 81 variables mix (Quantitatives et Qualitatives) qu'on peut les classer en fonction de leurs catégories :

Variables Démographiques, Caractéristiques de la Propriétés et du Zoning, Details sur le Terrain et les Environs, Qualité et état de la Maison, Details sur le toit et l'Extérieur, Details sur le sous-sol, Details sur le chauffage et l'Electricité, Surface des Etages, Nombre de Salles de bains, Details sur les Chambres et a Cuisine, Details fonctionnels et sur la Cheminée, Details sur le garage et l'Allee, Details sur les porches et les Terrasses, Piscine, Clôture et Caractéristique Diverses et finalement les détails des Ventes. Voici un tableau descriptif contenant des informations de structure sur notre Base de données :

Tableau 3: Tableau descriptive de la base de Données

Catégories	Nom	Description	Structure	Dimension
Démographique	Neighborhood	Localisation physique dans les limites de la ville d'Ames	Qualitative	0 NA
Propriété et Zonage	MSSubClass	Classe de bâtiment	Qualitative	0 NA
	MSZoning	Classification générale du zonage	Qualitative	4 NA
	Street	Type d'accès routier	Qualitative	0 NA
	Alley	Type d'accès à l'allée	Qualitative	2721 NA
	LotFrontage	Pieds linéaires de rue reliés à la propriété	Quantitative	486 NA
	LotArea	Superficie du terrain en pieds carrés	Quantitative	0 NA
	LotShape	Forme générale de la propriété	Qualitative	0 NA
	LandContour	Planéité de la propriété	Qualitative	0 NA
	Utilities	Type de services publics disponibles	Qualitative	2 NA
	LotConfig	Configuration du terrain	Qualitative	0 NA
	LandSlope	Pente de la propriété	Qualitative	0 NA
Terrain et Environs	Condition1	Proximité de la route principale ou du chemin de fer	Qualitative	0 NA
	Condition2	Proximité de la route principale ou du chemin de fer (si une seconde est présente)	Qualitative	0 NA
	BldgType	Type de logement	Qualitative	0 NA
	HouseStyle	Style de la maison	Qualitative	0 NA
Qualité et Etat de la Maison	OverallQual	Qualité générale des matériaux et de la finition	Qualitative	0 NA
	OverallCond	État général	Qualitative	0 NA
	YearBuilt	Date de construction initiale	Quantitative	0 NA
	YearRemodAdd	Date de rénovation	Quantitative	0 NA
	ExterQual	Qualité des matériaux extérieurs	Qualitative	0 NA
	ExterCond	État actuel des matériaux extérieurs	Qualitative	0 NA
	Foundation	Type de fondation	Qualitative	0 NA

Details sur le toit et l'Extérieur	RoofStyle	Type de toit	Qualitative	0 NA
	RoofMatl	Matériau du toit	Qualitative	0 NA
	Exterior1st	Revêtement extérieur principal	Qualitative	1 NA
	Exterior2nd	Revêtement extérieur secondaire (s'il y a plus d'un matériau)	Qualitative	1 NA
	MasVnrType	Type de placage en maçonnerie	Qualitative	24 NA
	MasVnrArea	Surface du placage en maçonnerie en pieds carrés	Quantitative	23 NA
Details sur le Sous-sol	BsmtQual	Hauteur du sous-sol	Qualitative	81 NA
	BsmtCond	État général du sous-sol	Qualitative	82 NA
	BsmtExposure	Murs de sous-sol avec sortie ou niveau jardin	Qualitative	82 NA
	BsmtFinType1	Qualité de la zone finie du sous-sol	Qualitative	79 NA
	BsmtFinSF1	Surface finie de type 1 en pieds carrés	Quantitative	1 NA
	BsmtFinType2	Qualité de la deuxième zone finie (si présente)	Qualitative	80 NA
	BsmtFinSF2	Surface finie de type 2 en pieds carrés	Quantitative	1 NA
	BsmtUnfSF	Surface non finie du sous-sol en pieds carrés	Quantitative	1 NA
	TotalBsmtSF	Surface totale du sous-sol en pieds carrés	Quantitative	1 NA
Details sur le Chauffage et l'Electricité	Heating	Type de chauffage	Qualitative	0 NA
	HeatingQC	Qualité et état du chauffage	Qualitative	0 NA
	CentralAir	Climatisation centrale	Qualitative	0 NA
	Electrical	Système électrique	Qualitative	0 NA
Surface des Étages	1stFlrSF	Surface du premier étage en pieds carrés	Quantitative	0 NA
	2ndFlrSF	Surface du deuxième étage en pieds carrés	Quantitative	0 NA
	LowQualFinSF	Surface finie de basse qualité (tous les étages)	Quantitative	0 NA
	GrLivArea	Surface habitable hors sol en pieds carrés	Quantitative	0 NA
Nombre de Salles de Bains	BsmtFullBath	Salles de bain complètes au sous-sol	Quantitative	2 NA
	BsmtHalfBath	Salles de bain semi-complètes au sous-sol	Quantitative	2 NA
	FullBath	Salles de bain complètes hors sol	Quantitative	0 NA
	HalfBath	Salles de bain semi-complètes hors sol	Quantitative	0 NA
Détails sur les Chambres et la Cuisine	BedroomAbvGr	Nombre de chambres au-dessus du sous-sol	Quantitative	0 NA
	KitchenAbvGr	Nombre de cuisines	Quantitative	0 NA
	KitchenQual	Qualité de la cuisine	Qualitative	1 NA
	TotRmsAbvGrd	Nombre total de pièces au-dessus du sol (hors salles de bain)	Quantitative	0 NA
	Functional	Évaluation de la fonctionnalité de la maison	Qualitative	2 NA

Détails Fonctionnels et sur la Cheminée	Fireplaces	Nombre de cheminées	Quantitative	0 NA
	FireplaceQu	Qualité de la cheminée	Qualitative	1420 NA
Détails sur le Garage et l'Allée	GarageType	Localisation du garage	Qualitative	157 NA
	GarageYrBlt	Année de construction du garage	Quantitative	159 NA
	GarageFinish	Finition intérieure du garage	Qualitative	159 NA
	GarageCars	Taille du garage en capacité de voitures	Quantitative	1 NA
	GarageArea	Surface du garage en pieds carrés	Quantitative	1 NA
	GarageQual	Qualité du garage	Qualitative	159 NA
	GarageCond	État du garage	Qualitative	159 NA
	PavedDrive	Allée pavée	Qualitative	0 NA
Détails sur les Porches et les Terrasses	WoodDeckSF	Surface de la terrasse en bois en pieds carrés	Quantitative	0 NA
	OpenPorchSF	Surface du porche ouvert en pieds carrés	Quantitative	0 NA
	EnclosedPorch	Surface du porche fermé en pieds carrés	Quantitative	0 NA
	3SsnPorch	Surface du porche trois saisons en pieds carrés	Quantitative	0 NA
	ScreenPorch	Surface du porche avec moustiquaire en pieds carrés	Quantitative	0 NA
Piscine, Clôture et Caractéristiques Diverses	PoolArea	Surface de la piscine en pieds carrés	Quantitative	0 NA
	PoolQC	Qualité de la piscine	Qualitative	2909 NA
	Fence	Qualité de la clôture	Qualitative	2348 NA
	MiscFeature	Caractéristique diverse non couverte dans d'autres catégories	Qualitative	2814 NA
	MiscVal	Valeur ($) de la caractéristique diverse	Quantitative	0 NA
Détails des Ventes	MoSold	Mois de la vente	Quantitative	0 NA
	YrSold	Année de la vente	Quantitative	0 NA
	SaleType	Le type de la vente (contrat ou cash…)	Qualitative	1 NA
	Salecondition	La condition de la vente (avec intérêt ou sans ...)	Qualitative	0 NA
	SalePrice	Le prix de chaque maison en $	Quantitative	1459 NA

Vous pouvez voir dans ce tableau que notre base de donnes contient des valeurs manquantes NA or, statistiquement parlant, ce n'est pas logique d'appliquer les modèles avant d'imputer ces valeurs manquantes et c'est pour cela on a remplacé les NA des variables quantitatives par leur moyenne et pour les variables qualitatives on les a remplacées par la modalité la plus fréquente pour chaque variable.

Voici une représentation graphique qui montre seulement les variables contenant des valeurs manquantes NA avec un pourcentage par rapport a la dimension des valeurs non manquantes. On remarque que les valeurs manquantes constituent 6.5% des valeurs de la base de donnes. Autrement dit, il existe 93.5% des valeurs dans notre base de données et 6.5% des NA.

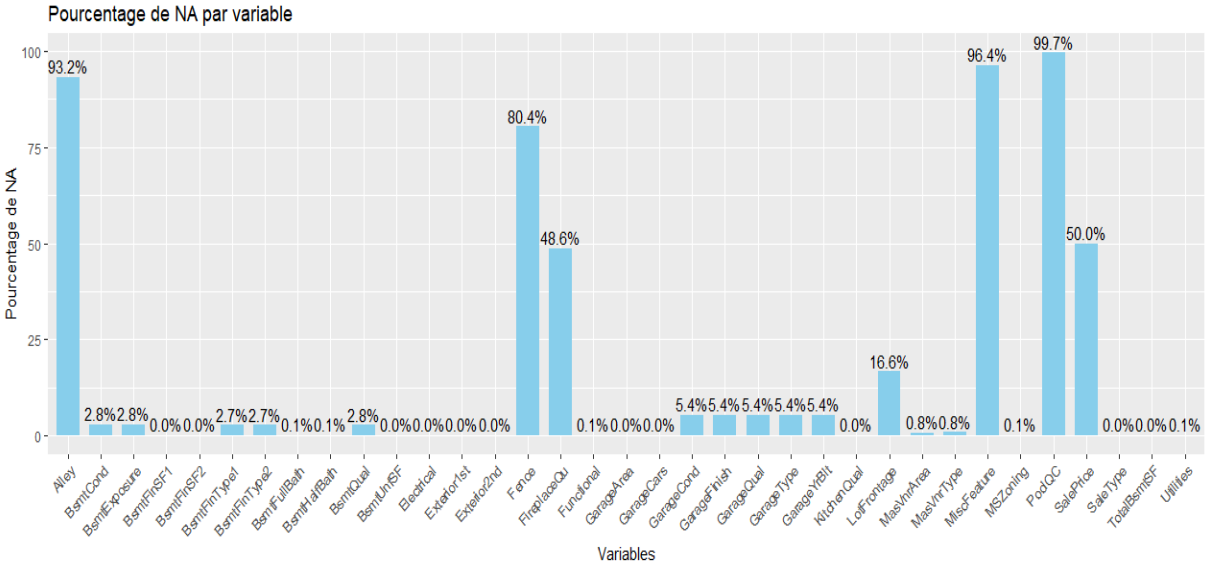

Figure 1: Représentation Graphique des pourcentages de NA par variables

1.2) Qu'est ce qu'on a cherché à évaluer ?

Notre étude vise à évaluer la performance de différents modèles de régression et de régularisations pour prédire les prix des maisons en se basant sur leurs caractéristiques. Mais avant de commencer à appliquer et évaluer il faut faire attention aux conditions (hypothèses) des modèles de régression qu'on va les démontrer avant de commencer à travailler :

Condition 1 : Linéarité : D'après les représentations graphiques de la variable cible prix aves les variables explicatives, on peut voir qu'il existe des variables qui ne sont pas linéaire avec la variable cible (SalePrice vs GarageArea) en outre, il existe aussi des variables explicatives qui sont linéaires (SalePrice vs YearBuilt & YearRemodADD), avec la variable cible ce qui montre une relation proportionnelle entre eux.

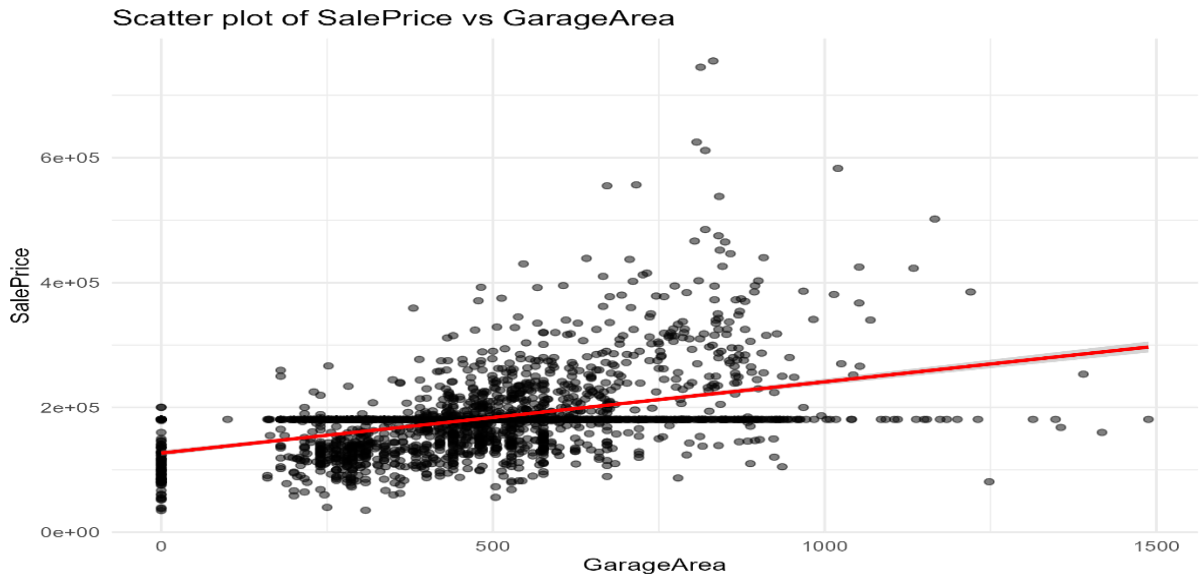

Figure 2: Représentation graphique de la variable cible Sale Price en fonction de la variable explicative Garage Area

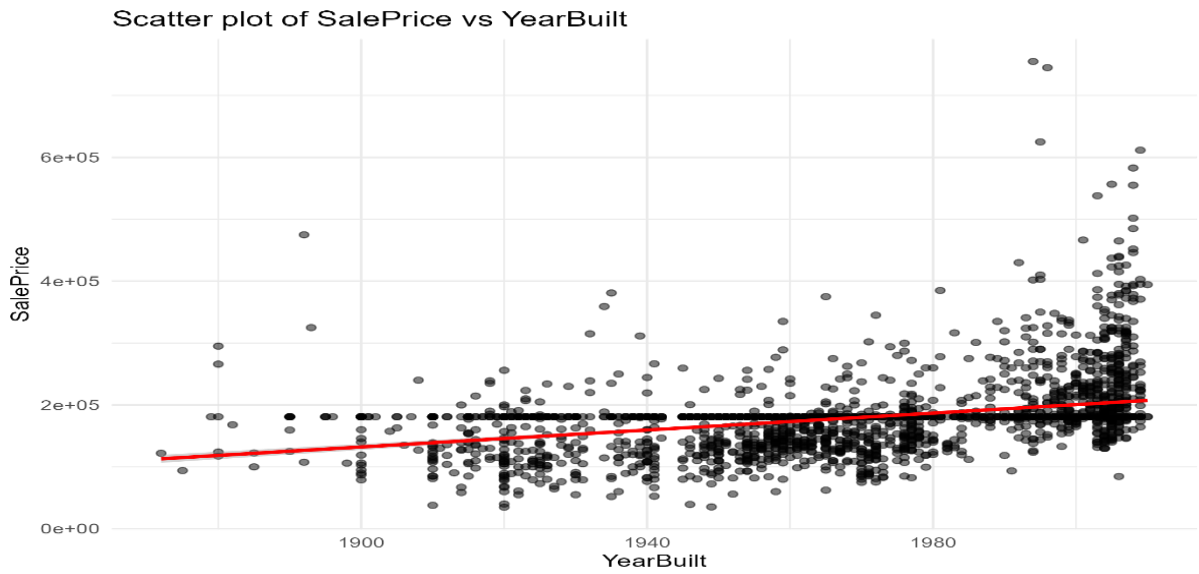

Figure 3: Représentation graphique de la variable cible Sale Price en fonction de la variable explicative Year Built

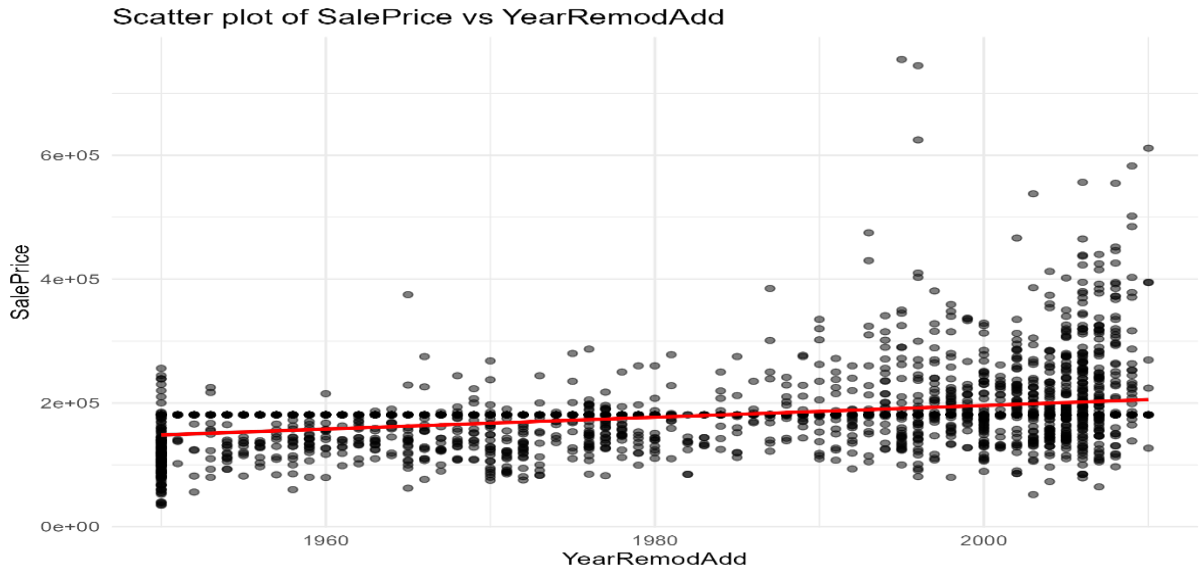

Figure 4: Représentation graphique de la variable cible Sale Price en fonction de la variable explicative Year Remod Add

Condition 2 : Homoscédasticité ou Hétéroscédasticité : Par définition l'homoscédasticité c'est lorsque la variance des erreurs de prédictions (les résidus) est constante pour toutes les valeurs des variables explicatives mais, si la variance des erreurs change avec le changement des valeurs des variables explicatives dans ce cas il existe une hétéroscédasticité. Pour vérifier l'homoscédasticité, on a tracé les résidus en fonction des valeurs prédites ou on a obtenu une dispersion homogène des erreurs.

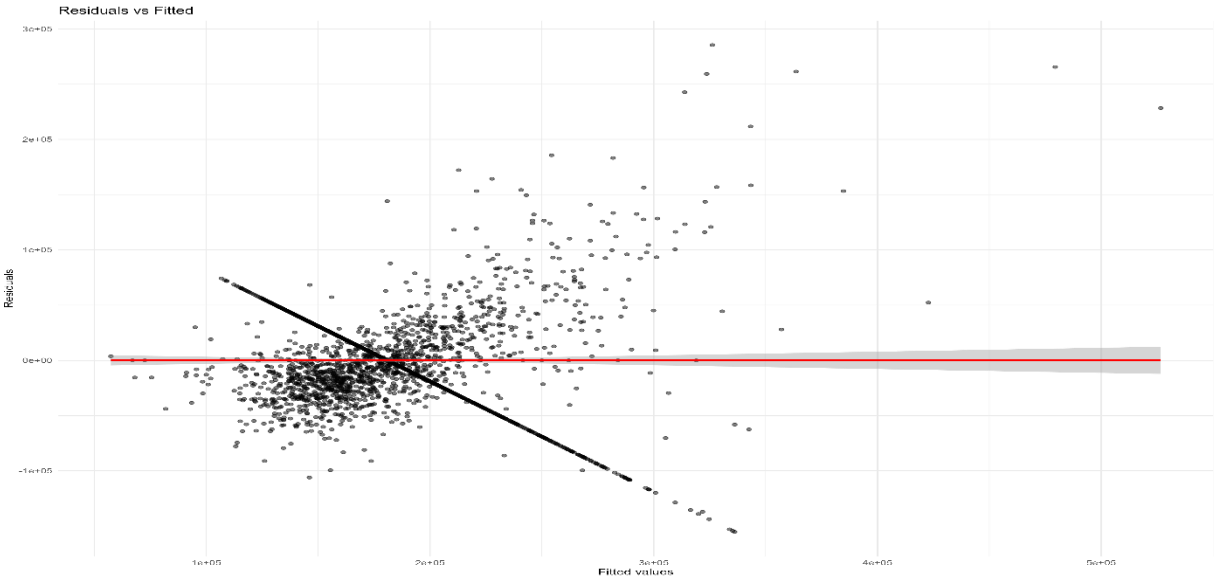

Figure 5: Représentation graphique des résidus en fonction des valeurs prédits

Condition 3 : Normalité : Cette condition est vérifiée lorsque les erreurs de prédictions suivent une distribution normale autrement dit, les valeurs des erreurs doivent être proche de zéro. On a montré la normalité d'après un histogramme des résidus

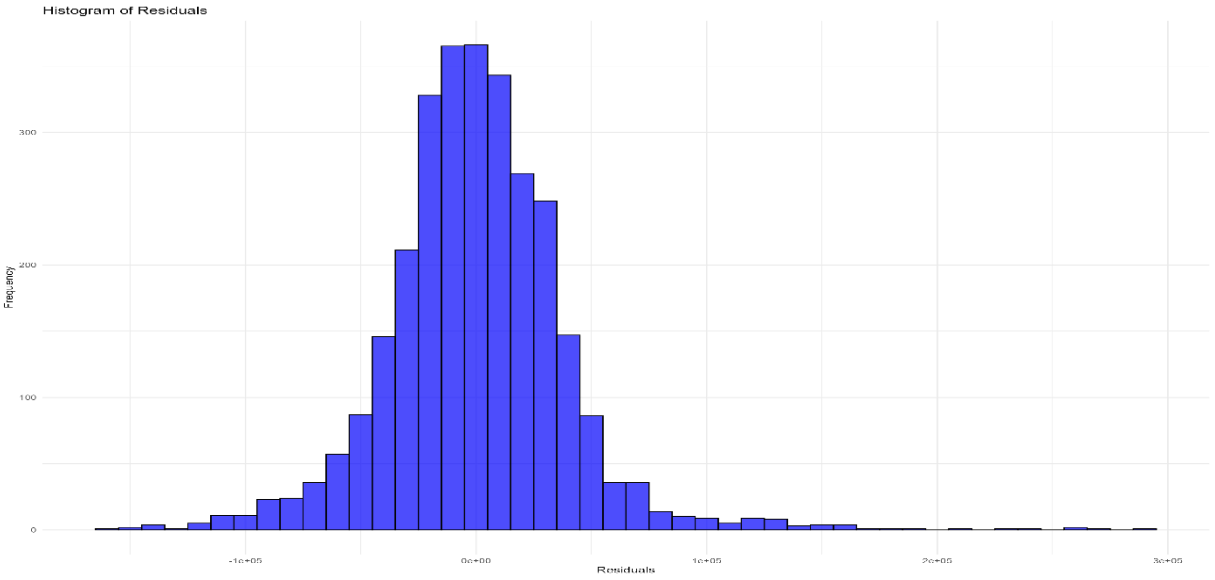

Figure 6: Histogramme montrant la distribution Normale des Résidus

Condition 4 : Multi colinéarité : La multi colinéarité existe lorsque 2 ou plusieurs variables explicatives sont fortement corrélées entre elles ce qui cause un surapprentissage et rend difficile l'évaluation de l'impact de chaque variable explicative sur le modèle. Voici la matrice de corrélation entre les 81 variables explicatives qui montre entre forte corrélation (couleur verte) entre des différentes variables explicatives.

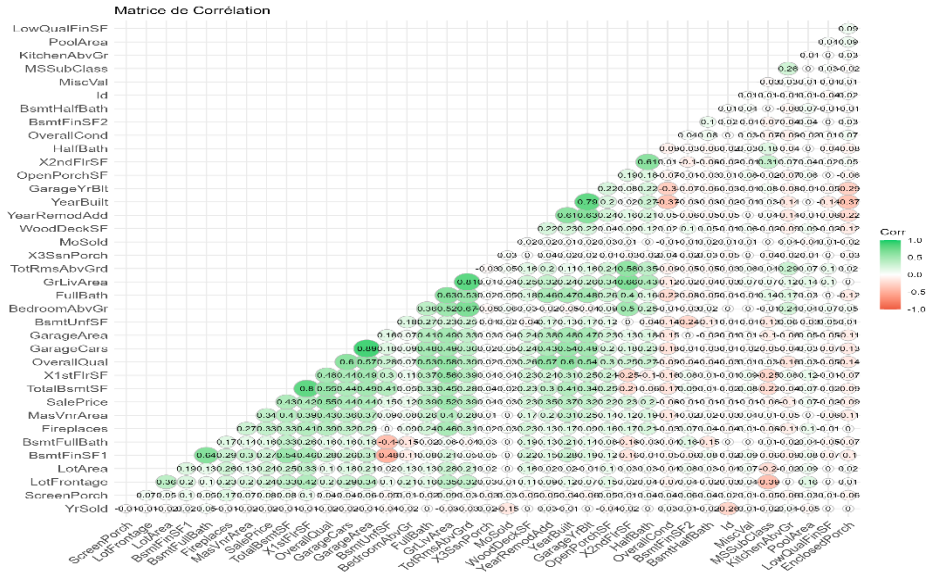

Figure 7: Matrice de Corrélation des variables explicatives

Transformation logarithmique, Codage et Partitionnement :

Dans cette partie on a essayé de transformer la variables cible en utilisant la fonction logarithmique c'est pour cela on a obtenu 2 cas d'évaluation **avant et après transformation logarithmique**. De plus, on a transformé notre base de données en des valeurs numériques (transformer les variables qualitatives en variables quantitatives) alors dans ce cas on aura 2 bases de données :

data_transformed_one_hot (ou on a appliqué le codage One hot qui transforme les variables qualitative en quantitative sans ordre et sans supposer des condition ou des hypothèse sur la transformation mais ce codage augmente le nombres des variables) et **data_tranformed_label** (ou on a appliqué le codage par Etiquette qui convertit les catégories des variables qualitatives en entiers, ce codage est simple et rapide mais peut être problématique pour les variables qualitatives non nominal).

Maintenant, avant de commencer par l'application des 7 modèles sur les 2 base de données on doit la diviser en deux partie (sous ensemble d'entraînement qui présente 80% des observations et sous ensemble de test qui présente 30% des observations).

Par définition, les modèles de régression linéaire sont des méthodes statistiques pour créer des combinaisons linéaires entre une variable cible (quantitative ou qualitative) et une/plusieurs variables explicatives (quantitatives ou qualitatives). Maintenant on va voir une explication théorique sur les modèles de régression avancés utilisés dans cette étude pour prédire les prix immobiliers :

Modèle 1 : OLS Régression

Régression des moindres carres ordinaires (Ordinary Least square Regression) est un des modèles de régression linéaire qui **minimise la somme des carrées des différences entre les valeurs réelles et les valeurs prédites** (sans éliminer aucune variable explicative) en donnant a la fin des valeurs d'erreurs de cette prédiction MSE, MAE, RMSE, R^2 et R^2 adj pour évaluer la performance de ce modèle.

Modèle 2 : Régression linéaire pas à pas (Ascendante)

Régression linéaire pas à pas (Ascendante) est un des modèles de régression linéaire qui **ajoute chaque fois une variable explicative** a la combinaison linéaire et test chaque fois l'erreur de prédiction de cette combinaison pour avoir a la fin la meilleur combinaison entre les variables explicatives choisies. A la fin il calcule les valeurs d'erreurs de cette prédiction MSE, MAE, RMSE, R^2 et R^2 adj pour évaluer la performance de ce modèle.

Modèle 3 : Régression linéaire sur Composantes Principales (PCR)

Régression linéaire sur Composantes Principales (PCR) est un des modèles de régression linéaire qui utilise les composantes principales pour réduire la dimensionnalité des variables explicatives et cela en transformant les variables explicatives en un ensemble des nouvelles variables non corrélées appelées composantes principales. C'est-à-dire PCR est une méthode **mix entre ACP et Régression linéaire** qui nous donne à la fin des valeurs d'erreurs de cette prédiction MSE, MAE, RMSE, R^2 et R^2 adj pour évaluer la performance de ce modèle.

Modèle 4 : Régression des moindres carres partiels (PLS)

Régression des moindres carres partiels (PLS) est un des modèles de régression linéaire qui optimise la covariance entre les variables explicatives et la variable cible. Cette méthode est **similaire à PCR** car elle réduit la dimensionnalité en projetant les variables explicatives et cibles dans un nouvel espace ou la covariance entre les variables est maximisées ce qui nous aide à traiter la multi colinéarité. A la fin, elle nous donne des valeurs d'erreurs de cette prédiction MSE, MAE, RMSE, R^2 et R^2 adj pour évaluer la performance de ce modèle.

Modèle 5 : Régression de Ridge

Régression de Ridge est une technique de régularisation appliquée lorsqu'il existe une multi colinéarités (comme dans notre base de données ou les variables sont fortement corrélées). Cette technique réduit le nombre des variables les moins importantes **en appliquant une pénalité L2** contrôlée pas le paramètre alpha, qui diminue l'influence des variables moins importantes ais ne les élimine pas. A la fin, elle nous donne des valeurs d'erreurs de cette prédiction MSE, MAE, RMSE, R^2 et R^2 adj pour évaluer la performance de ce modèle.

Modèle 6 : Régression de LASSO (LASSO)

Régression LASSO est une technique de régularisation appliquée lorsqu'il existe une multi colinéarités (comme dans notre base de données ou les variables sont fortement corrélées). Cette technique peut annuler certains coefficients **en appliquant une pénalité L1** contrôlée pas le paramètre alpha, ce qui facilite la sélection des variables significatives. A la fin, elle nous donne des valeurs d'erreurs de cette prédiction MSE, MAE, RMSE, R^2 et R^2 adj pour évaluer la performance de ce modèle.

Modèle 7 : Régression d'Elastic Net (Elastic Net)

Elastic Net est une technique de régularisation appliquée lorsqu'il existe une multi colinéarités (comme dans notre base de données ou les variables sont fortement corrélées). Cette technique peut a la fois annuler certains coefficients **en appliquant une pénalité L1** et réduire le nombre des variables les moins importantes en appliquant une pénalité L2 contrôlée par les paramètre alpha et beta.

1.3) Quelles sont les critères de jugements ?

Les résultats de performance de ces modèles seront bases sur les mesures métriques suivantes : Root Mean Square Errer **(RMSE**), coefficient de détermination **(R^2)** et l'Air sous la courbe **(AUC)** pour mesure la précision et la qualité de prédictions de ces modèles.

2. Résultats obtenus

Apres applications des 7 modèles de régression sur les deux bases de données (data_tranformed_on-hot et data_transformed_label), on a obtenu les différentes mesures métriques sur lesquelles nous allons baser notre étude pour sélectionner le meilleur modèle de prédiction des prix immobiliers.

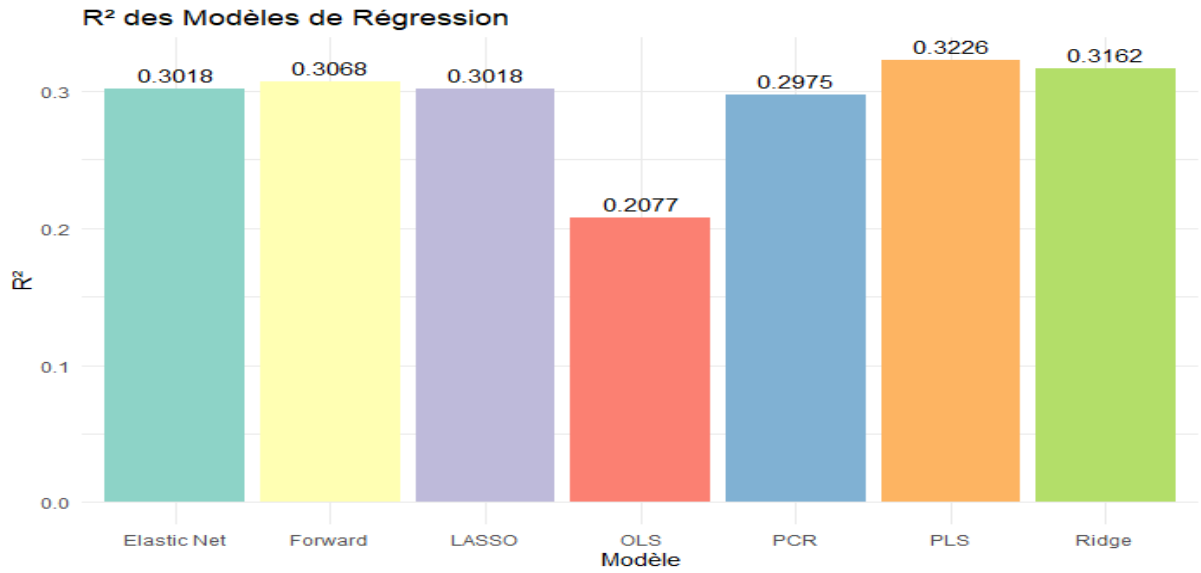

Figure 8: Graphe montrant les valeurs R^2 des 7 modèles pour data_transformed_label

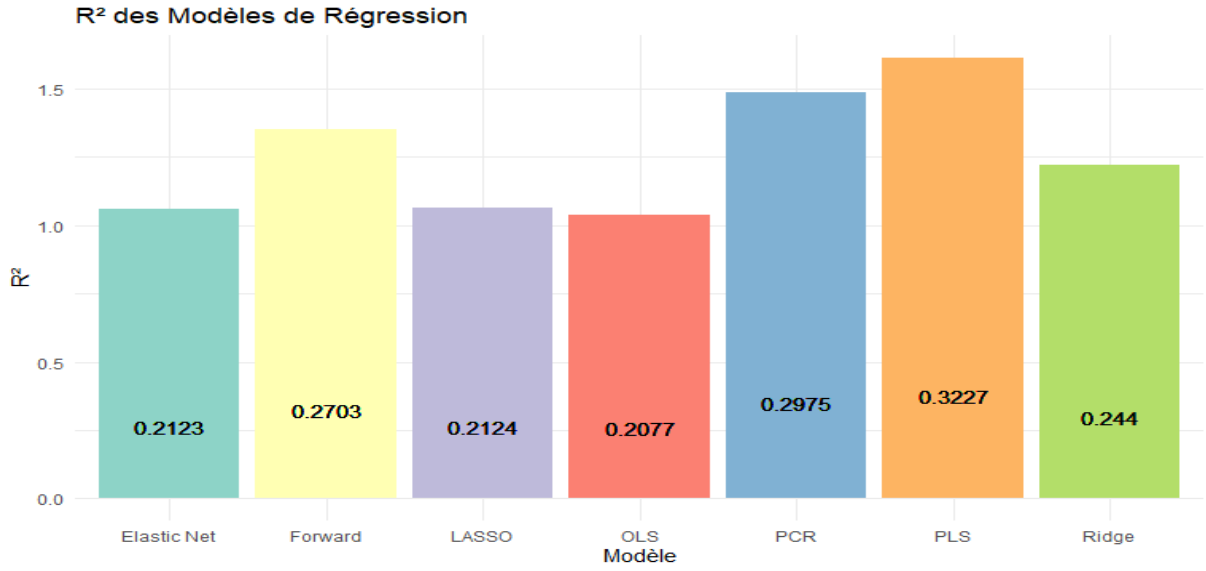

Figure 9: Graphe montrant les valeurs R^2 des 7 modèles pour data_transformed_one_hot

Comme vous savez les valeurs R^2 varient entre 0 et 1. si la valeur est proche de 1 cela signifie que le modèle est bien expliqué et si la valeur est proche de 0 cela signifie que le modèle n'explique pas bien les données. Alors qu'ici d'après les résultats obtenus on constate que toutes les modèles de régressions n'expliquent pas beaucoup notre base de données après transformation en utilisant les deux méthodes One-Hot coding et Label coding car l'ajustement ne dépasse pas le **30% - 33%.**

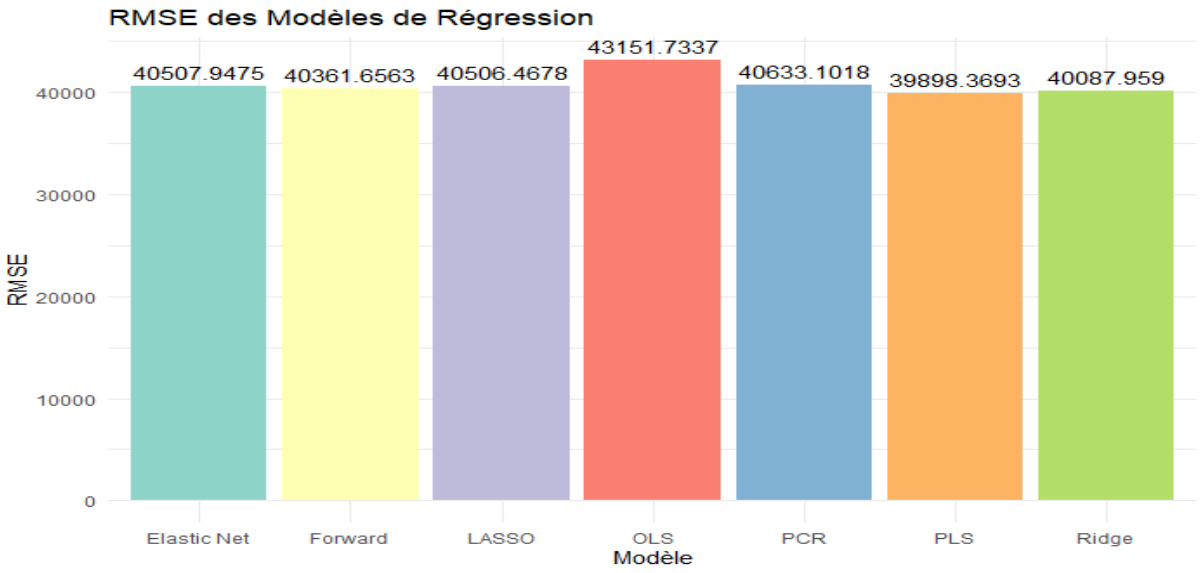

Figure 10: Graphe montrant les valeurs RMSE des 7 modèles pour data_transformed_label

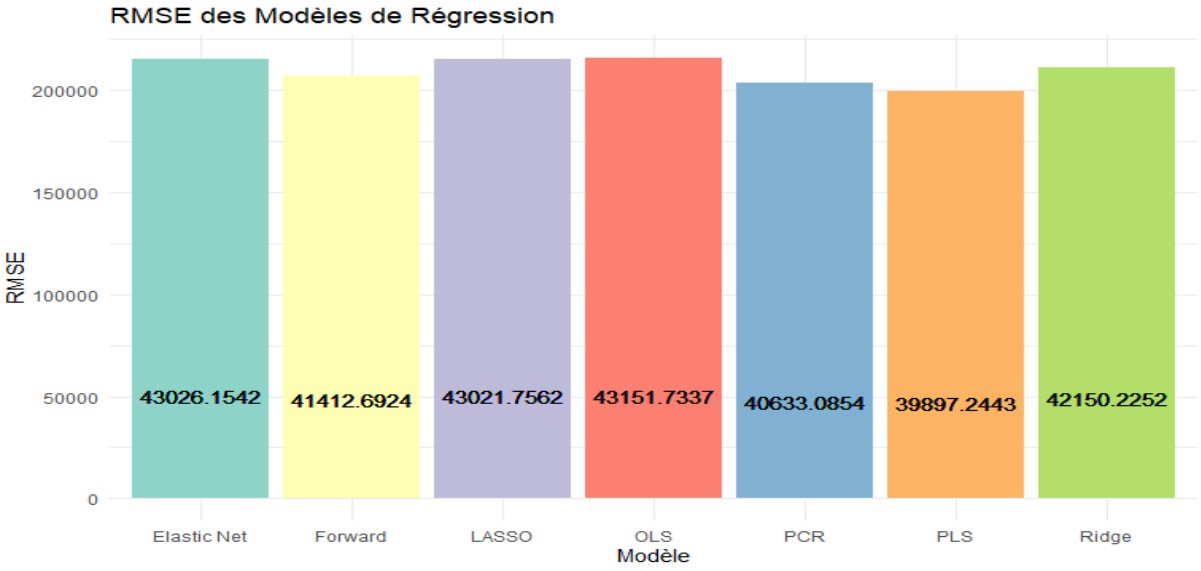

Figure 11: Graphe montrant les valeurs RMSE des 7 modèles pour data_transformed_one-hot

Comme vous savez aussi pour le RMSE, si la valeur est faible cela signifie que les valeurs prédites sont proches des valeurs observées (un meilleur ajustement du modèle) et si la valeur est élevée cela signifie que les valeurs prédites sont éloignées des valeurs observées (un mauvais ajustement). Alors qu'ici d'après les résultats obtenus la performance des prédictions des 7 modèles est très proche qui varient entre **39000 - 43000** car tous ont une RMSE moyenne (il n'existe pas un grand décalage entre les RMSE)

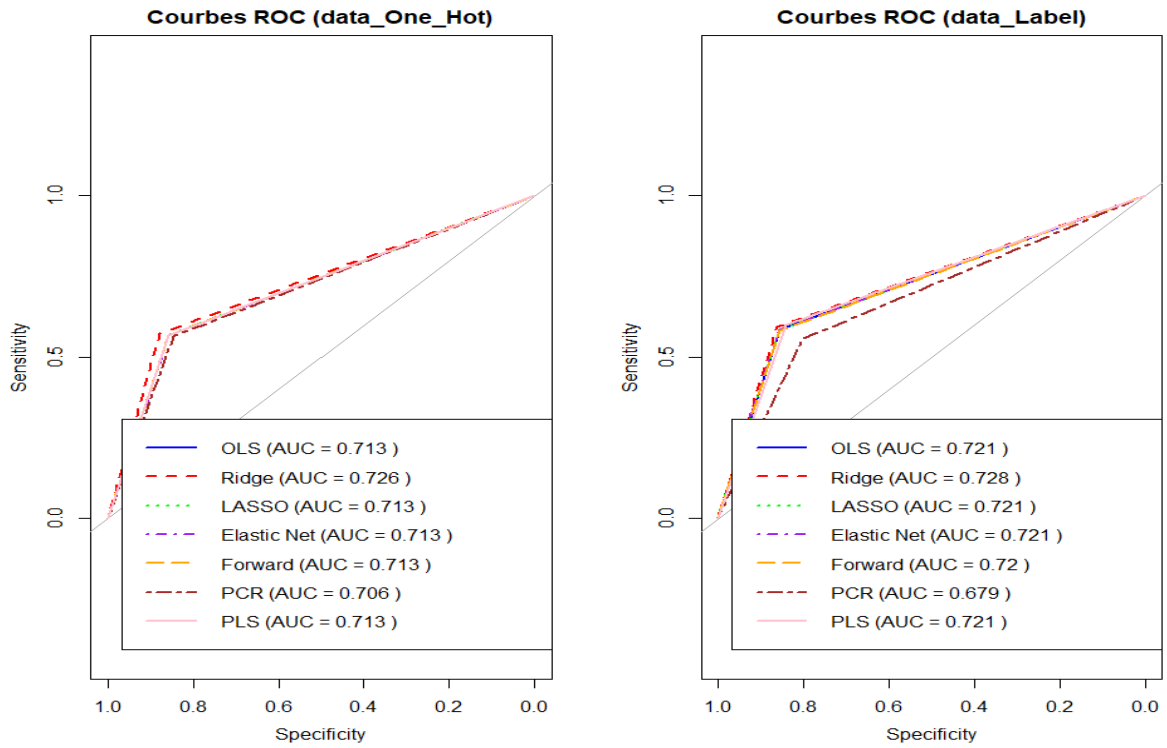

Figure 12: Courbes Rocs des deux bases de données transformées

Voici la Courbe Roc des deux bases de données transformées montrant les différentes valeurs des AUC qui sont très proches l'une des autres ce qui nous indique que les modèles ont une performance moyenne dans les prédictions car les valeurs varient entre **67% - 72%.** De plus on peut voir qu'il n'existe pas une grande différence entre la transformation One-Hot coding et Label coding

En se basant sur ces résultats, on a appliqué une **transformation logarithmique** de la variable cible dans le but de rendre la distribution plus normale et réduire l'impact des extremums. De plus on a testé si les mesures métriques vont changer, on a remarqué que les valeurs des RMSE ont diminuées ce qui nous montre que la performance des modèles est maintenant beaucoup mieux qu'avant pour faire la prédiction des prix immobiliers et choisir le meilleur modèle.

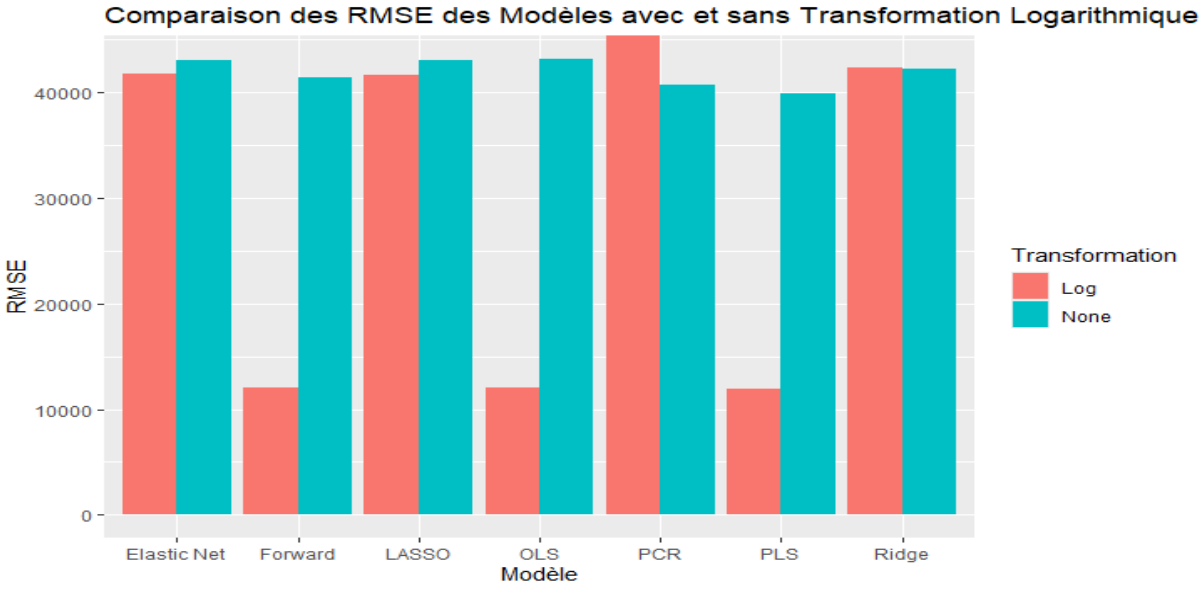

Figure 13: Représentation graphique montrant RMSE avant et après transformation logarithmique

IX. Conclusion

1. Rappel sur la problématique et les résultats obtenus

Rappelons que le problème de notre étude c'est d'aider les investisseurs à acheter ou vendre les maisons dans le meilleur prix en utilisant les différents modèles de régression. Alors après transformation logarithmique on a déduit que les 3 meilleurs modèles de régression sont : Régression linéaire pas à pas (**Ascendante**), Régression des moindres carres ordinaires (**OLS**) et Régression des moindres carres partiels (**PLS**) comme ayant les plus petites valeurs de RMSE alors que maintenant on peut voir d'après les représentation graphique que les deux modèles **OLS** et **Forward** s'executent plus rapide que le modèle **PLS** dans les deux bases de données **data_transformed_label** et **data_transformed_one_hot**.

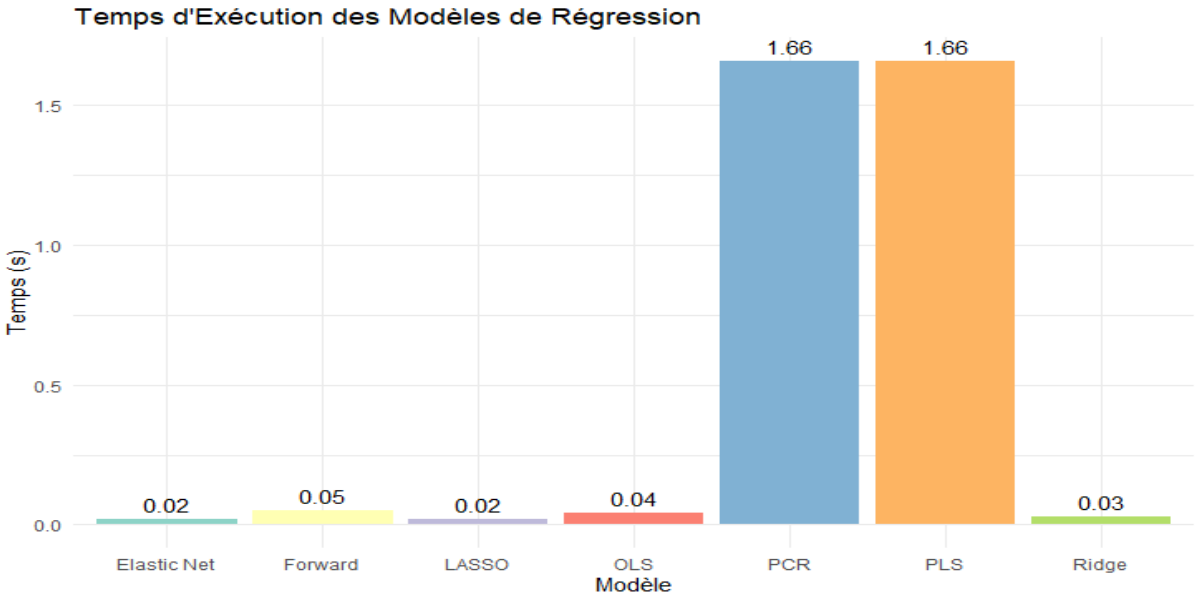

Figure 14: Temps d'exécution des 7 modèles pour data_transformed_label

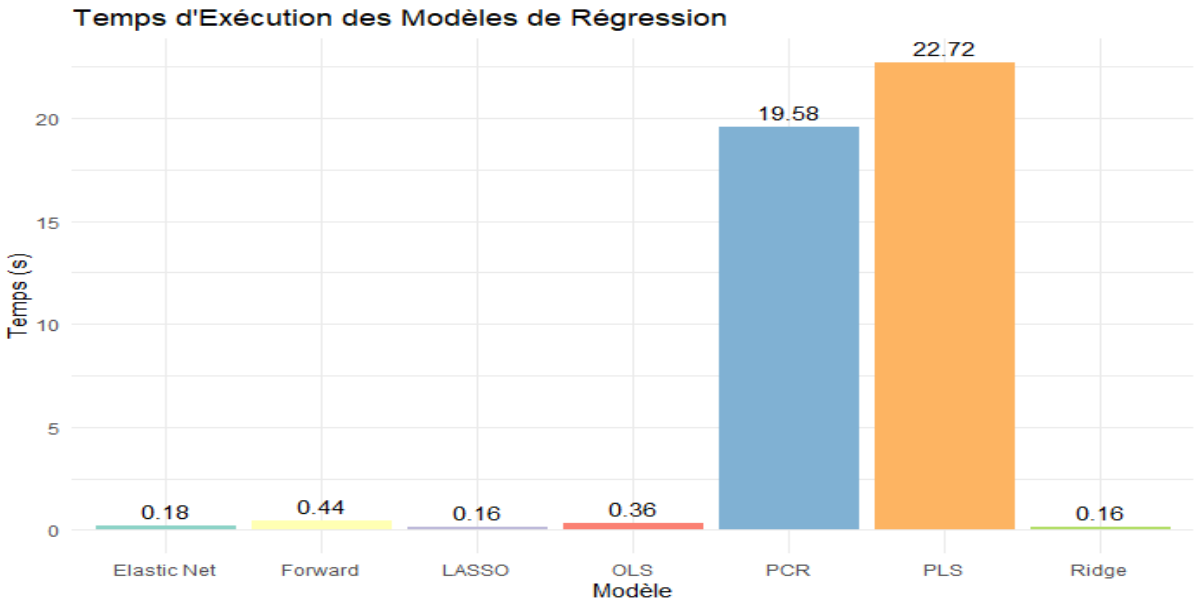

Figure 15: Temps d'exécution des 7 modèles pour data_transformed_one_hot

En résume, les 3 modèles de régressions : Ascendante, OLS et PLS ont une valeur **AUC=71%** dans la base de données data-transformed_label alors que dans la base de données data_transformed_one_hot ces 3 modèles ont une valeur **AUC=72%** ce qui montre qu'il n'y a pas beaucoup de différence entre les transformations et que la prédiction est acceptable mais on peut trouver d'autre modèles (complexes) de prédictions qui peuvent avoir des valeurs AUC plus grandes.

2. Les limites des recherches

En se basant sur les résultats de notre étude on constate que la base de données contient des NA et avant de commencer l'entrainement des 7 modèles on a imputé par la moyenne (valeurs numériques) et par la modalités la plus fréquentes (valeurs non numérique) de plus le codage one-Hot et Label étaient nécessaire pour l'application des 7 modèles ce qui affectent négativement sur les résultats (mesures métriques) de la base de données obtenus. Autrement dit, les valeurs des AUC n'ont pas dépassé 72% au lieu d'avoir des AUC entre 85%-95% c'est pour cela on conclut qu'on peut appliquer des modèles plus complexe (Random Forest, Boosting Bagging ou Réseaux de neurones ou autre modèles sophistiquées) sur cette base de données.

3. Une ouverture

Cette étude nous aide a prédire les prix immobiliers mais est ce qu'on peut collecter une base de données qui nous aide a prédire les prix des voiture et nous permette d'ouvrir une concessionnaire de voitures (marchand de voitures) ?

X. Les Recommandations

On se basant sur les résultats des mesures métrique, on recommande d'appliquer des modèles plus sophistiqués comme les modèles ensemblistes et les Réseaux de neurones.

XI. Bibliographie

https://www.kaggle.com/code/apapiu/regularized-linear-models
https://www.kaggle.com/code/pmarcelino/comprehensive-data-exploration-with-python
https://www.kaggle.com/code/masumrumi/a-detailed-regression-guide-with-house-pricing#Fitting-model(simple-approach)
https://www.kaggle.com/code/gusthema/house-prices-prediction-using-tfdf
https://www.kaggle.com/code/pmarcelino/data-analysis-and-feature-extraction-with-python
https://www.kaggle.com/c/house-prices-advanced-regression-techniques
https://rpubs.com/HishamElAdel/1102461
https://odr.inrae.fr/intranet/carto/cartowiki/index.php/Regression_lin%C3%A9aire_avec_R