

# MODÉLISATION DES PRIX DE L'IMMOBILIER À BOSTON AVEC EXTRATREESREGRESSOR



# Objectif et Données

Dataset contenant des informations sur différents quartiers de Boston, y compris le taux de criminalité, le pourcentage de zone résidentielle, le taux d'impôt foncier, etc.

**Le but de ce projet est de créer un modèle d'IA pour prédire les prix de l'immobilier à Boston. En utilisant des données sur les caractéristiques des logements et des quartiers, nous cherchons à déterminer les facteurs qui influencent le prix médian des logements dans chaque quartier. L'objectif final est de rendre les logements plus accessibles en fixant des prix plafonds basés sur des critères justes et précis.**

Nom de la colonne	Description
crime_rate	Taux de criminalité par habitant par ville.
residential_zone_pct	Proportion de terrains résidentiels zonés pour des lots de plus de 25 000 pieds carrés.
business_acres_pct	Proportion d'acres commerciaux non commerciaux par ville.
charles_river_dummy	Variable fictive pour la rivière Charles (1 si la parcelle borde la rivière, 0 sinon).
nitric_oxides_concentration	Concentration en oxydes nitriques (parties par 10 millions).
average_rooms	Nombre moyen de pièces par logement.
age	Proportion de logements occupés par leur propriétaire construits avant 1940.
distances_to_employment_centres	Distances pondérées jusqu'à cinq centres d'emploi de Boston.
accessibility_to_highways	Indice d'accessibilité aux autoroutes radiales.
property_tax_rate	Taux d'imposition foncière à pleine valeur par 10 000 \$.
pupil_teacher_ratio	Ratio élèves-enseignant par ville.
black_population	$1000(B_k - 0.63)^2$ où $B_k$ est la proportion de personnes de couleur par ville.
lower_status_pct	Pourcentage de la population de statut inférieur.
median_home_value	Valeur médiane des logements occupés par leur propriétaire en milliers de dollars.

# ● Choix du modèle


J'ai décidé d'utiliser les modèles RandomForestRegressor et ExtraTreeRegressor pour notre problème de modélisation des prix de l'immobilier à Boston. Ces choix ont été motivés par les meilleurs scores que j'ai obtenus en utilisant la bibliothèque PyCaret, qui propose une approche complète pour l'entraînement, l'évaluation et la sélection des modèles. Dans la phase d'évaluation, je vais comparer les performances de ces deux modèles pour déterminer celui qui offre la meilleure prédiction des prix de l'immobilier.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
et	Extra Trees Regressor	2.2705	11.0857	3.2205	0.8625	0.1411	0.1114	0.0460
gbr	Gradient Boosting Regressor	2.3624	12.4470	3.3886	0.8446	0.1488	0.1172	0.0370
rf	Random Forest Regressor	2.4547	14.0296	3.5799	0.8245	0.1530	0.1209	0.0430
lightgbm	Light Gradient Boosting Machine	2.5375	14.4021	3.6426	0.8235	0.1619	0.1283	0.0330
ada	AdaBoost Regressor	2.8693	17.1562	4.0157	0.7862	0.1773	0.1453	0.0290
lr	Linear Regression	3.5558	26.0318	5.0186	0.6851	0.2616	0.1771	0.3090
ridge	Ridge Regression	3.5361	26.1816	5.0335	0.6826	0.2647	0.1766	0.0230
lar	Least Angle Regression	3.6273	26.6034	5.0667	0.6764	0.2680	0.1814	0.0260
br	Bayesian Ridge	3.6077	26.9329	5.1177	0.6753	0.2695	0.1783	0.0240
dt	Decision Tree Regressor	3.3304	26.9692	4.8857	0.6617	0.2085	0.1592	0.0260
et	Extra Trees Regressor	3.7924	29.2767	5.3454	0.6594	0.2742	0.1798	0.0230
lasso	Lasso Regression	3.8270	29.9007	5.3974	0.6520	0.2717	0.1818	0.0240
llar	Lasso Least Angle Regression	3.8270	29.9003	5.3974	0.6520	0.2717	0.1818	0.0250
huber	Huber Regressor	3.7401	31.5595	5.5053	0.6189	0.2781	0.1816	0.0300
knn	K Neighbors Regressor	4.8667	49.1512	6.9196	0.4265	0.2613	0.2230	0.0270
omp	Orthogonal Matching Pursuit	6.1902	70.7901	8.3588	0.1764	0.3344	0.3011	0.0320
dummy	Dummy Regressor	6.8472	88.0726	9.3124	-0.0086	0.3967	0.3703	0.0240
par	Passive Aggressive Regressor	10.3730	208.0726	12.3877	-1.5159	0.5274	0.5951	0.0260





# Sélection du modèle et des hyperparamètres



**Pour trouver le meilleur modèle et les meilleurs hyperparamètres, j'ai utilisé une méthode de recherche approfondie. En utilisant la technique de GridSearchCV, j'ai évalué plusieurs modèles et combiné différentes valeurs d'hyperparamètres pour chaque modèle. Après avoir évalué les performances de chaque modèle, j'ai constaté que l'ExtraTreeRegressor a donné les meilleures performances en termes de prédiction des prix de l'immobilier à Boston. Par conséquent, j'ai choisi d'utiliser l'ExtraTreeRegressor comme modèle principal pour notre projet.**



# PRÉTRAITEMENT DES DONNÉES :

## Création de variables

La variable '`distance_accessibility_ratio`' : Ratio entre les distances aux centres d'emploi et l'accessibilité aux autoroutes

La variable '`crime_rate_black_population_ratio`' : Ratio entre le taux de criminalité et la population noire

## Suppression de variables

Suppression de la variable '`charles_river_dummy`' : Cette variable n'était pas pertinente pour notre modèle

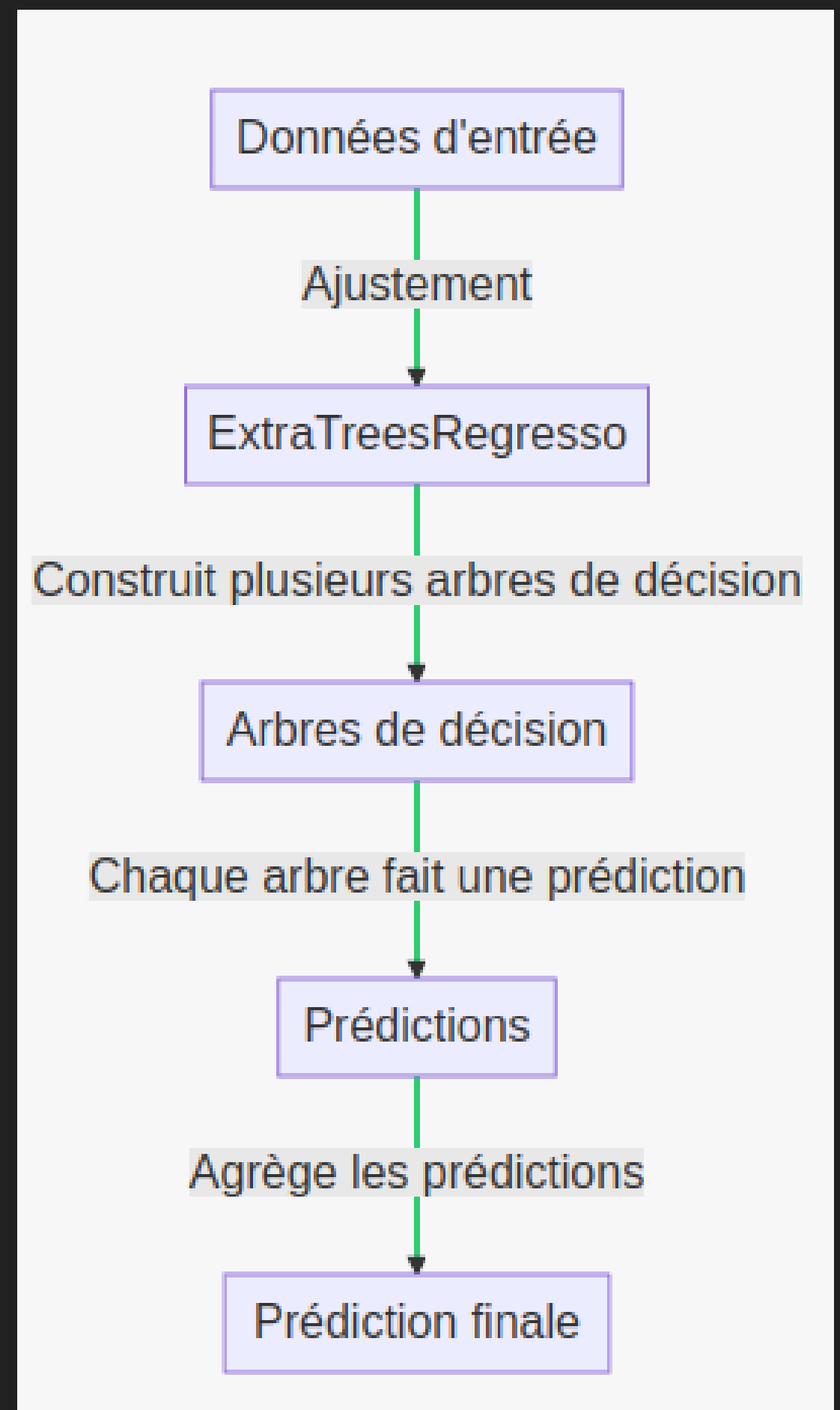
J'ai créé ces variables et supprimé '`charles_river_dummy`' pour améliorer la performance et la pertinence de notre modèle. La variable '`distance_accessibility_ratio`' prend en compte à la fois l'emplacement des emplois et l'accessibilité aux autoroutes, ce qui est important pour les acheteurs. Le ratio '`crime_rate_black_population_ratio`' nous permet d'explorer les relations entre la criminalité et la population noire dans chaque quartier, ce qui peut influencer les acheteurs. En supprimant '`charles_river_dummy`', nous simplifions notre analyse en éliminant une variable non pertinente.





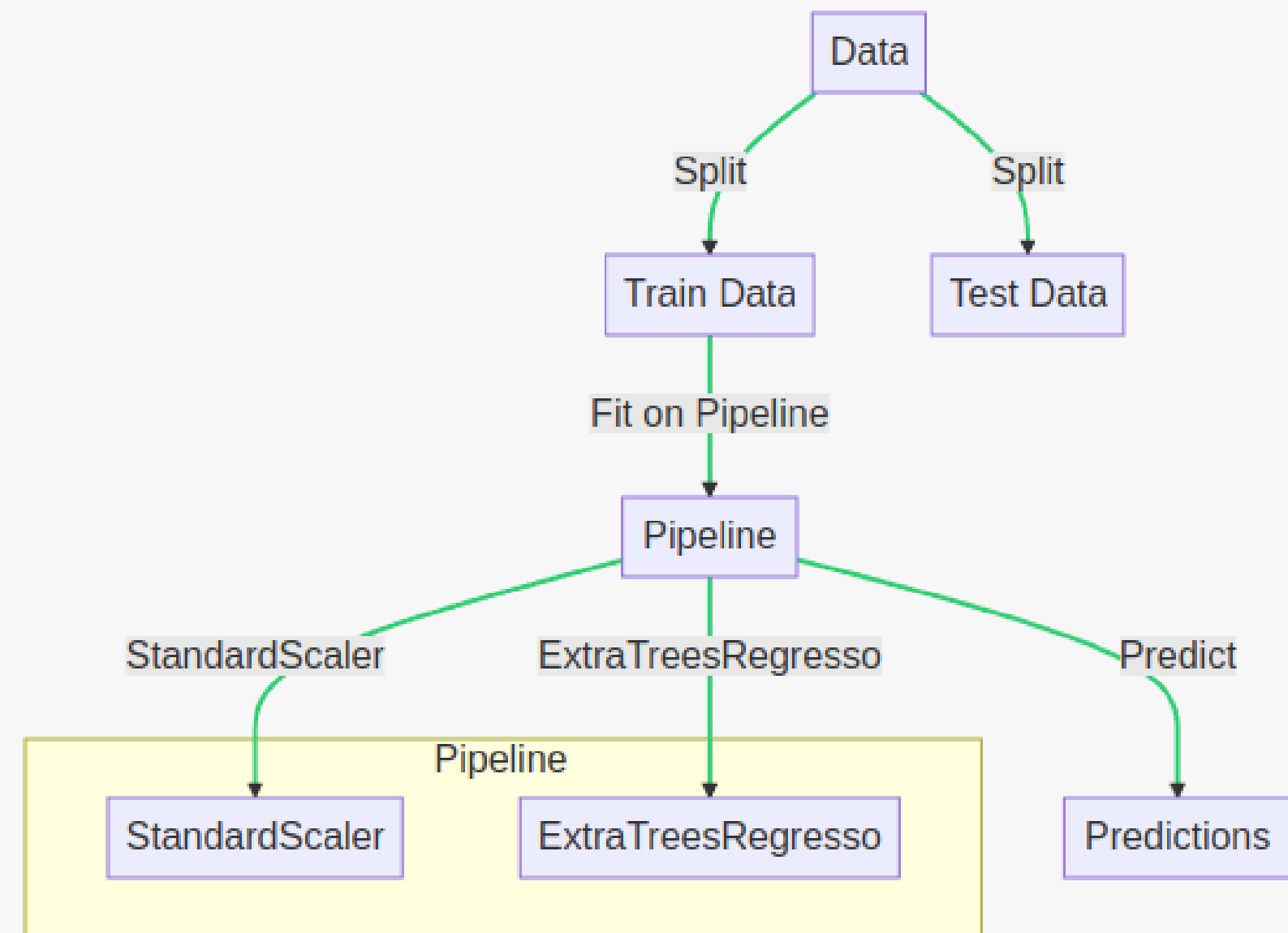
# MODÈLE

L'ExtraTreesRegressor est un algorithme de machine learning qui fonctionne en construisant un grand nombre d'arbres de décision sur divers sous-ensembles de données, puis en moyennant leurs prédictions pour obtenir une prédiction finale, ce qui le rend particulièrement efficace pour gérer les problèmes de surajustement et fournir une meilleure généralisation.



# Entraînement du modèle

Division des données : Les données ont été divisées en un ensemble d'entraînement (67% des données) et un ensemble de test (33% des données)



# ● Résultat

J'ai testé mon modèle sur différentes mesures d'évaluation, notamment le MAE (Mean Absolute Error), le RMSE (Root Mean Squared Error) et le R2 score. J'ai choisi de me concentrer principalement sur le RMSE.

Le RMSE est une mesure couramment utilisée pour évaluer l'exactitude des modèles de régression. Il accorde une plus grande importance aux erreurs importantes. Cela signifie que si notre modèle commet une grosse erreur sur une seule prédiction, le RMSE sera considérablement plus élevé. Dans notre cas, nous souhaitons minimiser les grosses erreurs, car elles peuvent avoir un impact significatif sur les prix de l'immobilier. Un RMSE plus faible indique que notre modèle commet moins d'erreurs importantes, ce qui est préférable pour résoudre notre problème de prédiction des prix de l'immobilier à Boston.

Mean Absolute Error: 1.946

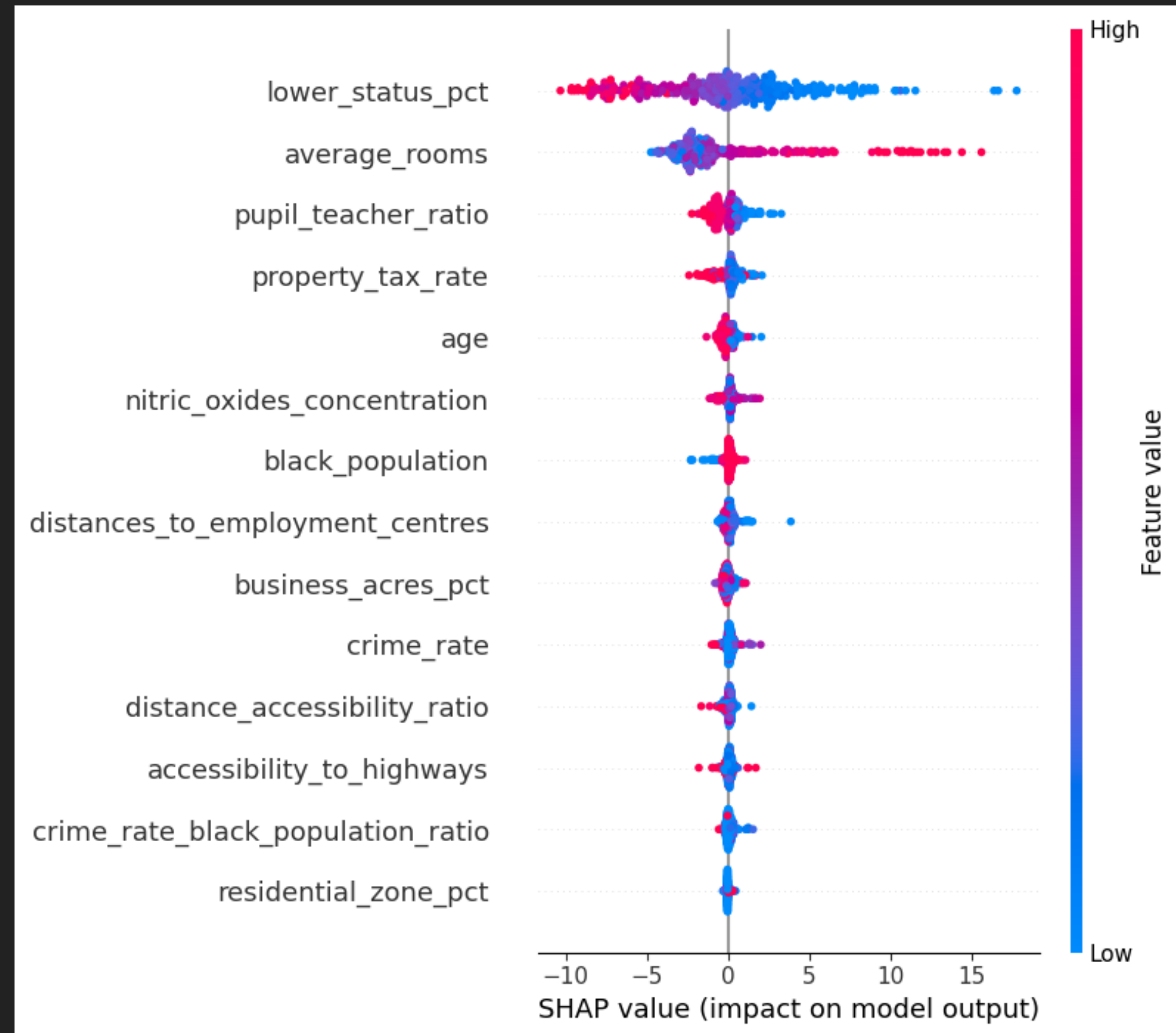
Root Mean Squared Error: 2.849

R2 Score: 0.893



# Importance des caractéristiques

Dans notre modèle de modélisation des prix de l'immobilier à Boston, deux caractéristiques se sont révélées particulièrement importantes : le "lower status percentage" et le nombre moyen de pièces ("average rooms"). Le pourcentage de statut inférieur a un impact négatif sur les prix, tandis que le nombre moyen de pièces a un impact positif. Ainsi, plus le pourcentage de statut inférieur est élevé, plus les prix ont tendance à baisser, tandis qu'une augmentation du nombre moyen de pièces est associée à une hausse des prix. Ces informations sont essentielles pour comprendre les facteurs qui influent sur les prix de l'immobilier à Boston.



# Conclusion

**En utilisant le modèle ExtraTreesRegressor et en sélectionnant les caractéristiques pertinentes, nous avons pu obtenir de bonnes performances de prédiction. L'analyse des valeurs SHAP nous a permis de comprendre l'importance des différentes caractéristiques et leur impact sur les prix de l'immobilier. Ces informations sont précieuses pour les professionnels du secteur immobilier, leur permettant de prendre des décisions éclairées et de mieux comprendre les facteurs qui influent sur les prix. Ce projet ouvre également la voie à d'autres possibilités d'amélioration et de développement de modèles plus sophistiqués pour une meilleure compréhension du marché immobilier.**