Rapport Scientifique – Prédiction des Prix Immobiliers à Boston

Machine Learning • LightGBM • Modélisation • MLflow • API FastAPI

---

1. Introduction

Ce projet vise à concevoir un modèle d'intelligence artificielle capable de prédire les prix de l'immobilier à Boston à partir d'un ensemble de caractéristiques socio-économiques, démographiques et environnementales.
Il s'inscrit dans un contexte d'aide à la décision pour **FastIA**, et s'appuie sur une démarche scientifique rigoureuse incluant :

- Analyse exploratoire complète
- Préparation et nettoyage des données
- Conception d'un pipeline reproductible
- Expérimentation de plusieurs modèles de régression
- Validation croisée et comparaison statistique
- Sélection du meilleur modèle
- Suivi automatique via MLflow
- Exposition du modèle via une API FastAPI

Ce README constitue un **rapport scientifique complet**, destiné à accompagner le notebook du projet.

---

2. Contexte métier – FastIA

FastIA souhaite développer un outil pédagogique et analytique permettant :

- d'illustrer la valeur ajoutée de l'IA dans l'immobilier ;
- de montrer comment les facteurs socio-économiques influencent le prix ;
- de disposer d'un pipeline ML clair, traçable et reproductible ;
- de tester le déploiement d'un modèle simple via API.

Le but **n'est pas** de produire un modèle utilisable en production, mais d'intégrer les bonnes pratiques ML.

---

3. Description du dataset Boston Housing

Le dataset contient **506 lignes** et **13 variables explicatives**, et une cible :

- **MEDV** : prix médian des logements (en milliers de dollars)

3.1 Variables explicatives

Variable	Description
CRIM	Taux de criminalité
ZN	Terrain résidentiel autorisé (zoning)
INDUS	Proportion d'activité non commerçante
CHAS	Proximité rivière Charles (binaire)
NOX	Polluants atmosphériques (NOx)
RM	Nombre moyen de pièces
AGE	Proportion d'anciens logements
DIS	Distance aux centres d'emploi
RAD	Accessibilité autoroutes
TAX	Taxe foncière
PTRATIO	Ratio élèves/professeurs
B	Indice ethnique (héritage statistique contestable)
LSTAT	% population défavorisée

---

4. Hypothèses scientifiques

1. Le prix augmente avec la taille du logement (**RM**).
2. Le prix diminue dans les zones défavorisées (**LSTAT**).
3. Les relations ne sont pas entièrement linéaires.

4. Les polluants (**NOX**) influencent négativement le prix.
5. Certains modèles non linéaires seront plus adaptés.

---

5. Analyse exploratoire (EDA)

Conclusions majeures de l'EDA :

- forte corrélation **RM ⟶ prix** ;
- forte corrélation négative **LSTAT ⟶ prix** ;
- relations non linéaires pour CRIM, NOX, TAX ;
- distribution du prix légèrement biaisée ;
- absence de valeurs manquantes ;
- outliers présents mais cohérents (logique urbaine).

➡️ Conclusion : un modèle non linéaire semble pertinent.

---

6. Préparation des données

6.1 Pipeline

Un pipeline scikit-learn a été construit pour assurer reproductibilité :

- **StandardScaler**
- transformation des colonnes via **ColumnTransformer**
- intégration avec le modèle ML (ex : LightGBM)

6.2 Split

- 80% entraînement
- 20% test
- random_state=42

➡️ Rien n'est fait "à la main" : tout passe par le pipeline.

---

7. Modèles testés

Trois modèles ont été évalués.

7.1 Régression Linéaire

- baseline
- interprétable
- mauvaise performance sur données non linéaires

7.2 RandomForest Regressor

- arbres de décision en ensemble
- robuste
- bon compromis biais/variance

7.3 LightGBM Regressor (sélectionné)

- modèle de gradient boosting
- très performant sur petits datasets
- capture très bien les non-linéarités
- hyperparamètres optimisés (learning rate, colsample, subsample…)

---

8. Méthodologie d'évaluation

8.1 Validation croisée (KFold = 5)

Pour garantir :

- robustesse ;
- stabilité ;
- faible variance de l'estimation des performances.

8.2 Métriques utilisées

- **RMSE** : racine erreur quadratique moyenne
- **MAE** : erreur absolue moyenne
- **R^2** : variance expliquée

---

9. Résultats & interprétation

Tableau des scores CV :

Modèle	RMSE	MAE	R^2
LightGBM	⭐ meilleur	⭐ meilleur	≈ **0.84**
RandomForest	bon	bon	≈ 0.83
Régression Linéaire	faible	faible	≈ 0.71

Analyse

- La régression linéaire ne capture pas les interactions.
- RandomForest fonctionne bien mais moins que LightGBM.
- LightGBM équilibre parfaitement biais et variance.

➡️ **LightGBM est retenu comme meilleur modèle.**

---

10. Analyse biais / variance

- **Linear Regression** : sous-ajustement
- **RandomForest** : faible variance, légèrement sur-ajusté
- **LightGBM** : optimum théorique pour ce dataset

---

11. Sauvegarde du modèle final

Le pipeline complet est sauvegardé sous : models/best_model.pkl

joblib.dump(best_pipe, "models/best_model.pkl")

---

12. Suivi expérimental via MLflow

MLflow a permis :

- suivi automatique des modèles,
- comparaison des métriques,

- sauvegarde des artefacts,
- traçabilité des hyperparamètres.

Interface :

mlflow ui --port 5000

---

13. Déploiement API – FastAPI

L'API expose :

13.1 Endpoint GET /

Retourne un message d'accueil.

13.2 Endpoint POST /predict

- entrée : JSON contenant les 13 variables
- sortie : prix prédit

Lancement :

uvicorn api.main:app --reload

---

14. Analyse éthique approfondie

14.1 Limites du dataset

- Dataset de 1978 → obsolète
- Contexte social biaisé
- Variable **B** issue d'un index racial → problématique
- LSTAT peut conduire à des discriminations indirectes

14.2 Risques pratiques

- modèle inadapté à une décision réelle
- risque de favoriser des inégalités sociales
- absence de modernité dans les variables

14.3 Mesures appliquées

- utilisation uniquement **pédagogique**
- documentation explicite des limites
- transparence de l'algorithme
- aucune recommandation d'usage commercial

---

15. Limites du projet

- Dataset trop petit (506 lignes)
- Dataset obsolète
- Peu de variables importantes absentes (surface, date, matériaux…)
- Valeurs manquantes minimes mais dataset très bruité
- Overfitting possible sur un dataset aussi réduit

---

16. Perspectives d'amélioration

- utilisation d'un dataset moderne (Kaggle real estate, Zillow, etc.)
- interprétabilité via SHAP
- ajout d'autres modèles (XGBoost, CatBoost)
- interface utilisateur Streamlit
- automatisation CI/CD du pipeline
- monitoring du drift en production

---

17. Conclusion générale

Le projet démontre :

- une démarche scientifique complète et reproductible
- un pipeline propre et modulable
- une validation rigoureuse
- un modèle final performant (LightGBM)

- un suivi des expériences via MLflow
- un déploiement grâce à FastAPI
- une prise en compte des enjeux éthiques et des limites

Le modèle n'a pas vocation à être utilisé en production, mais constitue une excellente démonstration de workflow machine learning complet.

---

18. Annexe – Quelques notions techniques

- RMSE = √(MSE)
- Cross-validation indispensable sur petit dataset
- StandardScaler ne doit être fit **que sur le train** → pipeline indispensable
- LightGBM utilise une approche histogram-based optimisée