21/10/2024

Analyse et Visualisation d'un Ensemble de Données sur un modèle de régression du prix de l'Immobilier.

Document rédigé Par le Groupe 1 :

- **Cesar Simo**
- **Samia**
- **Hiba**
- **Kevine**

Table des Matières

INTRODUCTION GÉNÉRALE

Dans un monde de plus en plus digitalisé, l'analyse de données joue un rôle crucial dans la prise de décision au sein des organisations. L'explosion des volumes de données générées quotidiennement a conduit les entreprises à intégrer l'analyse des données dans leurs stratégies, afin d'améliorer leur performance et de renforcer leur compétitivité. Cette discipline consiste à collecter, traiter, analyser et interpréter des informations afin d'en extraire des connaissances utiles pour orienter les décisions stratégiques.

L'analyse de données ne se limite pas à un simple traitement statistique : elle englobe des techniques avancées comme l'intelligence artificielle (IA), l'apprentissage automatique (machine Learning), ainsi que des approches plus classiques comme les analyses descriptives et prédictives. L'objectif est de transformer les données brutes en informations exploitables, permettant d'identifier des tendances, d'anticiper des comportements ou d'améliorer des processus internes.

Dans le cadre de ce travail, cette démarche d'analyse de données a été au cœur des missions confiées. En effet, les données peuvent provenir de la source (Kaggle), et son exploitation nécessitera des outils et méthodes adaptés. Parmi les étapes clés de l'analyse de données se trouvent :

1. **Collecte des données** : Acquisition et consolidation des données provenant de différentes sources.
2. **Nettoyage et préparation** : Traitement des données manquantes ou erronées afin d'assurer leur qualité.
3. **Exploration des données** : Visualisation et analyse préliminaire pour identifier des tendances ou anomalies.
4. **Modélisation et interprétation** : Application de modèles statistiques ou algorithmiques pour extraire des informations significatives.
5. **Communication des résultats** : Production de rapports et visualisations pour rendre les données compréhensibles par les décideurs.

Les outils utilisés pour l'analyse de données peuvent varier en fonction des besoins du projet, et inclure des plateformes comme Python (avec Pandas, NumPy, Seaborn), Power BI, Excel, ou encore des bases de données SQL.

Ce document présentera les différentes étapes de la faisabilité du projet notamment : **La description générale du sujet**, **les besoins fonctionnels et non fonctionnels**, **les besoins techniques**, **les contraintes et les livrables**.

PARTIE I : PRESENTATION GENERALE DU PROJET

Dans cette partie, nous allons en quelques mots décrire le projet en présentant le contexte général, les objectifs attendus et par ailleurs la délimitation fonctionnelle du sujet.

I. Présentation du Contexte

II. Problématique

III. Objectifs Globaux

IV. Délimitation du périmètre projet

I. PRÉSENTATION DU CONTEXTE

Dans le cadre des prévisions de la fluctuation des prix de l'immobilier sur le marché, un courtier immobilier, qui est notre client, souhaite obtenir une analyse des prix de l'immobilier afin d'améliorer ses décisions en matière de conseils. Pour cela, il existe plusieurs critères tels que la superficie, le nombre de chambres, la qualité du quartier, et autres qui permettront d'identifier la qualité d'une maison ainsi que son prix de vente. Indépendamment de ces facteurs, il existe un comportement du prix des maisons qui, pour notre client reste instable et par conséquent peu prévisionnel.

II. PROBLEMATIQUE

Dans un contexte de fluctuations des prix de l'immobilier, notre client, un courtier immobilier, cherche à affiner ses prévisions et à améliorer ses recommandations. La question centrale est la suivante : **Comment pouvons-nous identifier les données les plus pertinentes pour entraîner un modèle de régression capable de prédire avec précision le prix de l'immobilier à un moment donné ?**

III. OBJECTIFS GLOBAUX

Nous allons dans le cadre de notre travail implémenter un modèle de régression qui améliorera les prises de décisions et les conseils de notre client. Et celui-ci passera par des objectifs séquentiels telles que :

- **Visualisations des données** : Proposer des graphiques (nuage de points, distribution des prix, etc.) pour illustrer et identifier l'impact des différentes variables sur le prix des maisons.
- **Simulations personnalisées** : Permettre au courtier de modifier les paramètres (superficie, chambres, etc.) à travers une **interface interactive** intégré et de voir instantanément l'impact sur le prix projeté.
- **Filtres personnalisés :** Offrir des filtres interactifs permettant au courtier d'entreprendre des prévisions complexes à partir des données choisies a un moment défini.
- **Prédiction des prix :** Utiliser un modèle de régression pour prédire le prix des maisons en fonction des variables ayant le plus l'influence sur les prix de maison à un moment donné.

IV. DELIMITATION DU PERIMETRE DU PROJET

Le périmètre de notre sujet, sera établi à l'aide de l'ensemble des données recueillies dans le cadre de notre travail d'une part, et de l'ensemble des questions à répondre qui constitueront les fonctionnalités de notre système. Dans ce paragraphe, nous nous limiterons à présenter l'ensemble des données obtenues de notre source. Parmi lesquelles :

- **Square_Footage** : Taille de la maison en pieds carrés.
- **Num_Bedrooms** : Nombre de chambres dans la maison.
- **Num_Bathrooms** : Nombre de salles de bains.
- **Year_Built** : Année de construction de la maison.
- **Lot_Size** : Taille du terrain en acres.
- **Garage_Size** : Capacité du garage (nombre de voitures).
- **Neighborhood_Quality** : Qualité du quartier (échelle de 1 à 10).
- **House Price** : Le prix de la maison (variable cible à prédire).

Ce jeu de données, et de la définition des besoins fonctionnelles présenter dans la suite du travail, seront notre boussole dans l'implémentation formelle de notre modèle.

PARTIE II : SPECIFICATIONS FONCTIONNELLES ET NON FONCTIONNELLES

Cette partie, mettra en évidence l'ensembles des fonctionnalités dont notre modèle qui sera en charge de faire. Elle prendra en compte les exigences émises par le client et celles émises par le modèle.

I. Spécifications Fonctionnelles

1. Fonctionnalités Principales et Descriptions
2. Présentation de la Solution

II. Spécifications non Fonctionnelles

1. Gestion de l'Ergonomie et de la Performance
2. Gestion des Utilisateurs et des droits d'accès

I. SPECIFICATIONS FONCTIONNELLES

1. Fonctionnalités Principales et Descriptions

La délimitation fonctionnelle de notre modèle s'appuiera sur un ensemble de questions à répondre parmi lesquelles :

- Quels facteurs influencent le plus le prix des maisons ?
  - **Description:** Quelles variables parmi la taille de la maison, le nombre de chambres, de salles de bains, l'année de construction, la taille du terrain, la qualité du quartier et la taille du garage ont un impact significatif sur le prix des maisons ?
- Comment peut-on prédire efficacement le prix d'une maison ?
  - **Description :** Quel modèle de régression permet de prédire avec précision le prix des maisons en fonction des caractéristiques disponibles ?
- Existe-t-il une relation linéaire entre les caractéristiques des maisons et leurs prix ?
  - **Description :** Les relations entre le prix des maisons et les variables explicatives telles que la taille de la maison, le nombre de chambres et de salles de bains sont-elles linéaires ?
- Comment la qualité du quartier affecte-t-elle le prix des maisons ?
  - **Description :** Dans quelle mesure la variable Neighborhood_Quality (qualité du quartier) influence-t-elle le prix des maisons par rapport à d'autres caractéristiques comme la taille de la maison ou du terrain ?
- L'âge des maisons a-t-il un impact sur leurs valeurs ?
  - **Description:** Les maisons plus anciennes sont-elles moins chères en raison de l'usure, ou y a-t-il d'autres facteurs compensateurs, comme des rénovations ou des emplacements privilégiés ?
- La taille du terrain et la taille de la maison sont-elles corrélées ?
  - **Description:** Les maisons avec des terrains plus grands sont-elles également plus grandes en termes de superficie habitable, ou existe-t-il une variabilité dans la relation entre la taille du terrain et la taille de la maison ?
- Y a-t-il des biais dans les données ?
  - **Description:** L'ensemble de données contient-il des biais, tels que des concentrations de maisons dans certains types de quartiers ou de tranches de prix, qui pourraient influencer l'analyse et la prédiction ?

2. Présentation de la Solution

Notre modèle de façon global permettra d'apporter des réponses importantes à l'ensemble des descriptions présentées plus haut entre autres notamment :

- Identifier les caractéristiques ayant **la plus grande influence** sur le prix afin de comprendre les principaux facteurs déterminants du marché immobilier.
- Développer un **modèle de prédiction fiable**, permettant au courtier (agent immobilier) de mieux évaluer le prix des biens immobiliers sur la base de leurs caractéristiques.
- Vérifier si une régression linéaire est le **modèle le plus adapté** ou si des relations plus complexes existent entre les variables.
- Comprendre si les maisons situées dans des quartiers de haute qualité se vendent systématiquement plus cher, indépendamment de la taille ou d'autres caractéristiques physiques.
- Comprendre si les maisons situées dans des quartiers de haute qualité se vendent systématiquement plus cher, indépendamment de la taille ou d'autres caractéristiques physiques.
- Examiner comment l'année de construction influence les prix des maisons et si cette relation est constante dans toutes les gammes de prix.
- Explorer la relation entre la taille du terrain et la taille de la maison pour déterminer si les deux facteurs sont liés et comment ils influencent le prix.
- Identifier et traiter d'éventuels biais dans l'échantillon pour garantir la robustesse des résultats.

II. SPECIFICATIONS NON-FONCTIONNELLES

1. Gestion de l'Ergonomie et de la Performance

Aspect crucial dans la conception de l'interface de notre modèle, l'Ergonomie participera à un usage confortable et efficace de la gestion de nos données à travers un interface utilisateur (UI), l'interface utilisateur (UX), et l'accessibilité.

Nous prendrons notamment en charge les éléments ergonomiques suivant dans l'implémentation de notre modèle :

- **Facilité d'utilisation**
- **Réduction des erreurs**
- **Accessibilité**
- **Efficacité**

Par ailleurs, **la performance** se réfèrera à la rapidité et à l'efficacité avec lesquelles notre modèle exécute ses fonctions, ce qui inclut le temps de réponse, la fluidité d'exécution, et la capacité à gérer un grand nombre de requêtes simultanées.

Nous prendrons notamment en charge les éléments de performances suivant dans l'implémentation de notre modèle :

- **Temps de réponse rapide**
- **Optimisation des ressources**
- **Scalabilité**
- **Stabilité**

Nous mettrons en place de façon définitive un modèle a la fois rapides, fiables, et faciles à utiliser.

2. Gestion des Utilisateurs et des Droits d'Accès

Dans cette partie nous nous attèlerons à comprendre et à définir qui peut accéder à quoi, comment, et dans quelles conditions, afin de protéger les données et les ressources tout en permettant aux utilisateurs de travailler efficacement. Ceci se fera par la définition des autorisations et ou privilèges attribués à un utilisateur.

Afin de matérialiser cela dans notre modèle nous prendrons en compte les éléments suivants :

- **Création des Comptes**
- **Authentification**
- **Surveillance et gestion des Comptes**
- **Suppression des comptes**

PARTIE III : BESOINS TECHNIQUES

Nous parlerons dans cette partie de l'ensemble des outils et technologies qui participeront à la réalisation de notre modèle

I. Technologies Utilisées

II. Architecture de Déploiement

III. Compatibilité et Sécurité

I. TECHNOLOGIES UTILISEES

Le tableau ci-dessous présente l'ensemble des technologies utilisées dans le cadre de notre travail.

Noms	Types d'outils	Versions	Portabilités (Oui/Non)
Python	Langage de Programmation	3.12.4 | packaged by Anaconda, Inc.	Oui
Anaconda	Distribution Open Source	Conda version : 4.8.3	Oui
Jupyter Notebook	Environnement de Développement Open-Source	RAS	Oui
KAGGLE	Sources de Données	RAS	Oui
Excel	Tableur de gestion de données de la suite Office.	Version 20	Oui
Pandas	Framework Python	RAS	Oui
Numpy	Framework Python	RAS	Oui
Seaborn et Matplotlib	Framework Python	RAS	Oui
Sklearn	Framework Python	RAS	Oui
Zoom, WhatsApp	Plateforme de communication interactive.	RAS	Oui
HTML, CSS	Outils de construction des pages web pour nos visualisations interactives.	Version 60	Oui

Tableau1 : Technologies Utilisées

II. ARCHITECTURE ET DÉPLOIEMENT

Tableau ci-dessous présente l'architecture et les ramifications entre les outils et Framework utilisés dans le cadre de notre travail.

Noms	Fonctions
1. **Framework Panda**	Bibliothèque utilisée pour charger et manipuler les données.
2. **Framework Seaborn, Matplotlib**	Bibliothèques utilisées pour visualiser et identifier les tendances, les modèles et les relations dans les données.
3. **Jupyter Notebook**	L'outil permettra entre autres de sauvegarder le fichier de travail sous forme de fichier**. ipynb** et est exporter en HTLM ou PDF.
4. **Google Colab**	Environnement Cloud open source, partager entre collaborateur du projet, où les fichiers de travail peuvent être exportés et ou créés.

5. Sklearn	Bibliothèque utilisée pour la machine Learning, il permettra de produire efficacement notre modèle de régression linéaire.
6. GitHub, Git	Environnement cloud de gestion du versioning de nos fichiers de travail et contrôle de gestion.
7. Google, Microsoft Edge	Navigateurs de déploiements de nos pages web.

Tableau 2 : Modèle d'Architecture et De Déploiement

III. COMPATIBILITÉ ET SÉCURITÉ

La compatibilité et la sécurité sont deux aspects importants dans l'élaboration des interfaces de visualisation. Voici un aperçu de chacun de ces concepts :

La **compatibilité** fait référence à la capacité d'un système, ou d'un modèle à fonctionner avec d'autres systèmes. Voici quelques dimensions de la compatibilité :

1. Compatibilité logicielle : Assure que les applications peuvent fonctionner sur différentes versions de systèmes d'exploitation ou avec d'autres applications. Par exemple, une application doit être compatible avec les différentes versions de Windows, MacOs ou Linux.
2. Compatibilité matérielle : Désigne la capacité d'un logiciel à fonctionner sur différents types de matériel (CPU, GPU, etc.) ou de périphériques (imprimantes, scanners, etc.).
3. Compatibilité inter-applications : Permet aux différentes applications de communiquer et d'échanger des données sans problème. Cela inclut l'utilisation de standards ouverts et d'API.

La **sécurité** concerne la protection des systèmes, des réseaux et des données contre les accès non autorisés, les cyberattaques et d'autres menaces. Voici quelques aspects de la sécurité :

1. Confidentialité : Protection des données sensibles pour qu'elles ne soient accessibles qu'aux personnes autorisées. Cela inclut des mesures comme le chiffrement et l'authentification.
2. Intégrité : Assure que les données ne sont pas altérées ou corrompues pendant le stockage ou la transmission. Des mécanismes comme les sommes de contrôle (checksums) ou les signatures numériques sont utilisés à cette fin.
3. Disponibilité : Garantit que les systèmes et les données sont accessibles aux utilisateurs autorisés lorsque cela est nécessaire. Cela implique des mesures de sauvegarde et de récupération après sinistre.
4. Authentification et autorisation : Les systèmes doivent s'assurer que les utilisateurs sont bien qui ils prétendent être (authentification) et qu'ils n'accèdent qu'aux ressources pour lesquelles ils ont des droits (autorisation).

PARTIE IV : CONTRAINTES

Dans cette partie, nous présenterons de nombreux aspects de notre projet lié au Calendrier du projet, aux Contraintes financières, et aux ressources Humaines et Matérielles.

I. Calendrier du projet

II. Contraintes Financières

III. Ressources Humaines et Matérielles

I. CALENDRIER DU PROJET

Le tableau suivant présente les différentes phases de découpage de notre projet.

Nom	Date de Début	Date de fin	Durée	Responsable
Analyse préalable, Identification et recueil des données.	2024/10/15	2024/10/25	08 Jours	• Chef de projet • Analystes • Client
Rédaction du cahier de charges Fonctionnel et appréciation.	2024/10/28	2024/11/15	12 Jours	• Chef de projet • Analystes
Rédaction du cahier de charges technique et appréciation.	2024/11/18	2024/11/22	05 Jours	• Chef de Projet
Implémentation du code et test.	2024/11/25	2024/12/13	15 Jours	• Chef de projet • Analystes
Déploiement, et validation.	2024/12/16	2024/12/17	02 Jours	• Chef de projet • Analystes
Rédaction du Guide Utilisateur.	2024/12/18	2024/12/18	01 Jour	• Chef de projet
Rédaction du cahier de visualisation.	2024/12/19	2024/12/20	02 Jours	• Chef de projet
Durée Estimée			**45 jours**	

Tableau 3 : Calendrier du Projet

Sur la base de notre tableau établi ci-dessus, notre projet débute effectivement le **2024/10/15** et s'achève le **2024/12/20**. Ainsi pour une durée estimative de **45 Jours**.

II. CONTRAINTES FINANCIERES

Les contraintes financières liées à notre travail engloberont les revenus de chaque intervenant par heure. Cette définition financière s'appuiera sur l'outil **TALENT** qui nous a permis de fixer les prix sereinement.

Postes	Prix/Heure	Nombres d'heures	Total ($)
Chef de projet	55$/Heure	180	9900
Analyste de Données	47.41$/Heure	120*3	17067
Développeur Front-end	35$/Heure	60	2100
Total :		360	**29067 $**

Tableau 4 : Estimation Financières

Suite a l'élaboration de notre estimation financière, le coût de total de ce projet en termes de main d'œuvre s'élèvera autour de : **Vingt-neuf mille soixante-sept dollars**.

III. RESSOURCES HUMAINES ET MATERIELLES

1. Ressources Humaines

Ce paragraphe nous présente les différents intervenants du projet et leurs principales fonctions dans l'organigramme de travail. Le tableau ci-après apporte une parfaite illustration :

Noms	Titres	Rôles	Commentaire
Cesar Simo	Chef de Projet	Organiser et gérer l'équipe de travail. Produire la documentation et les rapports d'analyses effectués. Intervenir dans les procédures d'analyses et de codages.	Le chef de projet est un membre à part entière de l'équipe de travail qui pour le cas d'espèce a une place de meneur dans le suivi et la finalisation du projet.
Samia	Analyste de données	Intervenir et produire les besoins fonctionnels du modèle. Intervenir dans la l'implémentation du code. Intervenir dans le recueil et la compréhension des données.	L'analyste de données est une personne au cœur de la réalisation du projet a travers ses interventions dans la compréhension, la conception et l'implémentation du modèle envisagé.
Hiba	Analyste de données	Intervenir et produire les besoins fonctionnels du modèle. Intervenir dans la l'implémentation du code. Intervenir dans le recueil et la compréhension des données.	L'analyste de données est une personne au cœur de la réalisation du projet à travers ses interventions dans la compréhension, la conception et l'implémentation du modèle envisagé.
Kevine	Analyste de données	Intervenir et produire les besoins fonctionnels du modèle. Intervenir dans la l'implémentation du code. Intervenir dans le recueil et la compréhension des données.	L'analyste de données est une personne au cœur de la réalisation du projet à travers ses interventions dans la compréhension, la conception et l'implémentation du modèle envisagé.
Client X	Partie Prenante	Intervenir dans les validations itératives de chaque niveau d'avancement du projet	Le Client est une entité ou ressource importante au cours de l'avancement et de la validation du projet final.

Tableau 5 : Ressource Humaines

2. Ressources Matérielles

Nous présenterons ici l'ensemble des ressources matérielles que nous utiliserons pour produire notre travail.

Nom	Fonction	Quantité
Laptop	Outil principal de travail	04
Système d'exploitation Windows	Système de gestion logiciel	04
Modem Wifi	Outil pour connexion internet	04

Tableau 5 : Ressources Matérielles

PARTIE V : LIVRABLES ATTENDUS

Cette partie de notre travail présentera en quelques point les livrables ou élément à fournir à la fin de notre travail.

I. Documents à Produire

II. Modules Livrables

I. DOCUMENTS A PRODUIRE

Il s'agit pour l'équipe de travail d'élaborer l'ensemble de la documentation lié au projet pour une meilleure compréhension d'une part et une meilleure traçabilité de gestion d'autres part. A cet effet, les documents suivants seront à remettre au client a la finalisation du projet :

- **Guide Utilisateur** : Ce Document qui décrira les procédures de gestion et de manipulation des différents interfaces produit pour le client.

- **Cahier des rapports de visualisation** : Ce Document présentera de façon succincte l'ensembles des graphes et explications liées aux différentes préoccupations du clients tel qu'énuméré dans la description fonctionnelle.

- **Le Rapport de Test et Validation** : Ce document comportera les différentes phases d'implémentations du code ainsi que les commentaires y afférentes et par ailleurs, les niveaux de validations effectués avec les parties prenantes.

II. MODULES LIVRABLES

Nous livrerons au client dans le cadre de notre projet les éléments suivants :

- **Une Interface Interactive** : Cette interface comportera les différents champs ou colonnes ou données qui permettront au client aisément de faire des simulations prédictives en fonction du jeu de données qu'il aura choisi.

- **Les Fichiers Sources** : ces jeux de fichiers seront constitués de l'ensemble du code source ayant permis l'implémentation du projet.

CONCLUSION GENERALE

L'analyse de données est aujourd'hui un **levier essentiel pour la performance des organisations**, leur permettant de prendre des décisions plus éclairées, d'améliorer leurs processus et d'anticiper les tendances du marché. Tout au long de mon stage, nous avons pu observer l'impact direct que peut avoir une exploitation pertinente des données sur l'efficacité des stratégies adoptées par l'entreprise. En mobilisant diverses techniques et outils d'analyse, ce projet a permis de **transformer des données brutes en informations exploitables**, renforçant ainsi la prise de décision.

En conclusion, cette expérience a été **enrichissante sur les plans technique et professionnel**. Elle a permis de consolider mes connaissances en analyse de données tout en renforçant la capacité à travailler en équipe et à répondre aux besoins d'une organisation. Convaincu que l'analyse de données représente un **outil indispensable** pour toute entreprise cherchant à innover et à rester compétitive.