

PROJET PYTHON / ML
M2 SISE

OCTOBRE 2023



À propos

Brentwood, Missouri

GreenTech Solutions est une société de service qui développe des applications.

Une agence immobilière sollicite l'entreprise afin d'aider ses commerciaux à mieux estimer les biens à vendre.

Pour cela vous devrez développer une application permettant de mieux comprendre le marché et d'estimer le prix de vente des nouveaux appartements.

Votre équipe

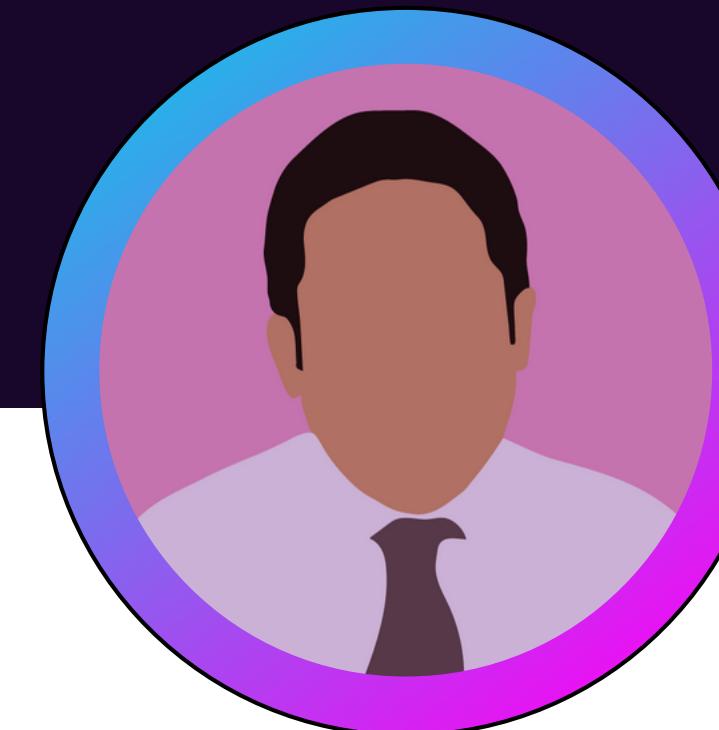
Pour ce projet, vous serez en équipe de 3 ou 4. Les équipes seront tirées au sort dans un même groupe de TP.



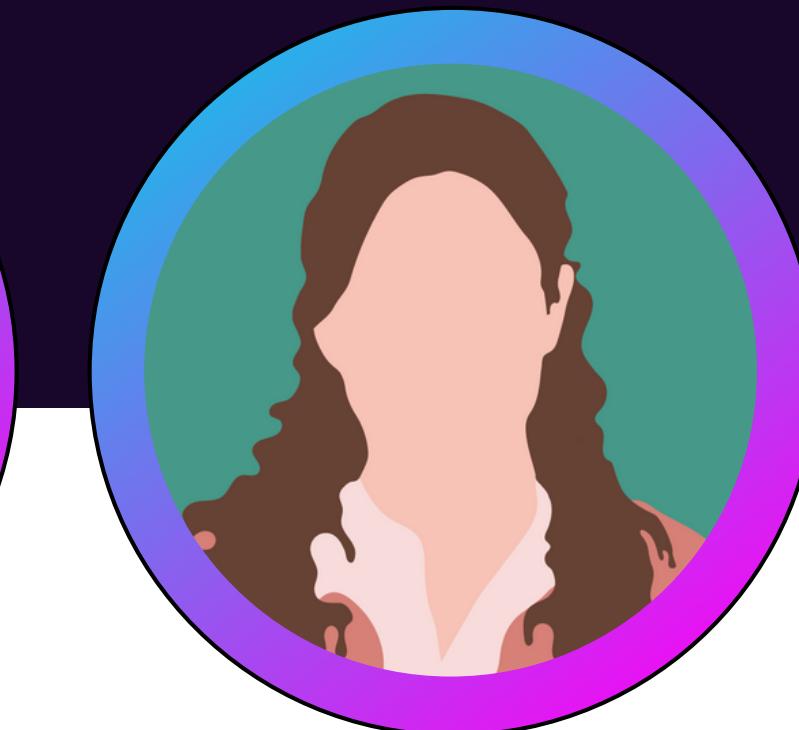
Anthony
Client



Student 1
Chef de projet



Student 2
Développeur



Student 3
Data Scientist



Les rôles

CHEF DE PROJET

est chargé de planifier, coordonner et gérer toutes les activités liées au projet afin de fournir une solution efficace, fonctionnelle et répondant aux besoins des utilisateurs.

DATA SCIENTIST

joue un rôle essentiel dans l'analyse des données, la préparation des données et la création de visualisations pertinentes afin de fournir des informations significatives et exploitables aux utilisateurs finaux. Il est également chargé de construire le modèle.

DÉVELOPPEUR

joue un rôle essentiel le développement et le déploiement de la solution. Il veille à effectuer des tests pour éviter les bugs. Il garantit le bon fonctionnement et la performance de la solution développée.



Les fonctionnalités



Standard

- Crédit d'un modèle de prédition de prix de vente
- Versionning et documentation sur GitHub
- Statistiques sur les ventes (KPI, tableaux, graphiques)
- Cartographie
- Filtres pour filtrer les statistiques
- Plusieurs onglets sur l'application
- Déploiement de l'application en local



Intermédiaire

- Fonctionnalités Standard
- Bouton pour exporter les graphiques en .png
- Une page "prédition" pour estimer un bien selon des paramètres
- Enrichir l'apprentissage du modèle avec des données d'OpenData
- Déploiement de l'application sur le web



Expert

- Fonctionnalités Intermédiaire
- Le user peut réentrainer le modèle en important des données
- Crédit d'une API (Application Programming Interface) pour votre modèle d'analyse prédictive



Les livrables

Pour ce projet, le client attend plusieurs livrables. Il ne se contentera pas de l'application. Il est important pour lui d'être autonome sur la maintenance de l'application une fois celle-ci opérationnelle.



01

Déploiement de l'application

L'application doit-être déployée sur <https://render.com/> OU <https://shiny.posit.co/py/> OU <https://www.heroku.com/students;>

02

Participation à la compétition Kaggle

Inscrivez votre équipe sur la compétition privée. L'objectif est de prédire le montant de la vente d'un bien.

03

Documentation technique et fonctionnelle

Cette partie du projet n'est pas à négliger. Elle comprend :

- le repos GitHub est documenté avec un README.md le code et les données brutes
- documentation technique (orientée pour développeur) avec schéma de l'architecture dessiné avec [Draw.io](#)
- documentation fonctionnelle (orientée utilisateur)
- rapport d'analyse

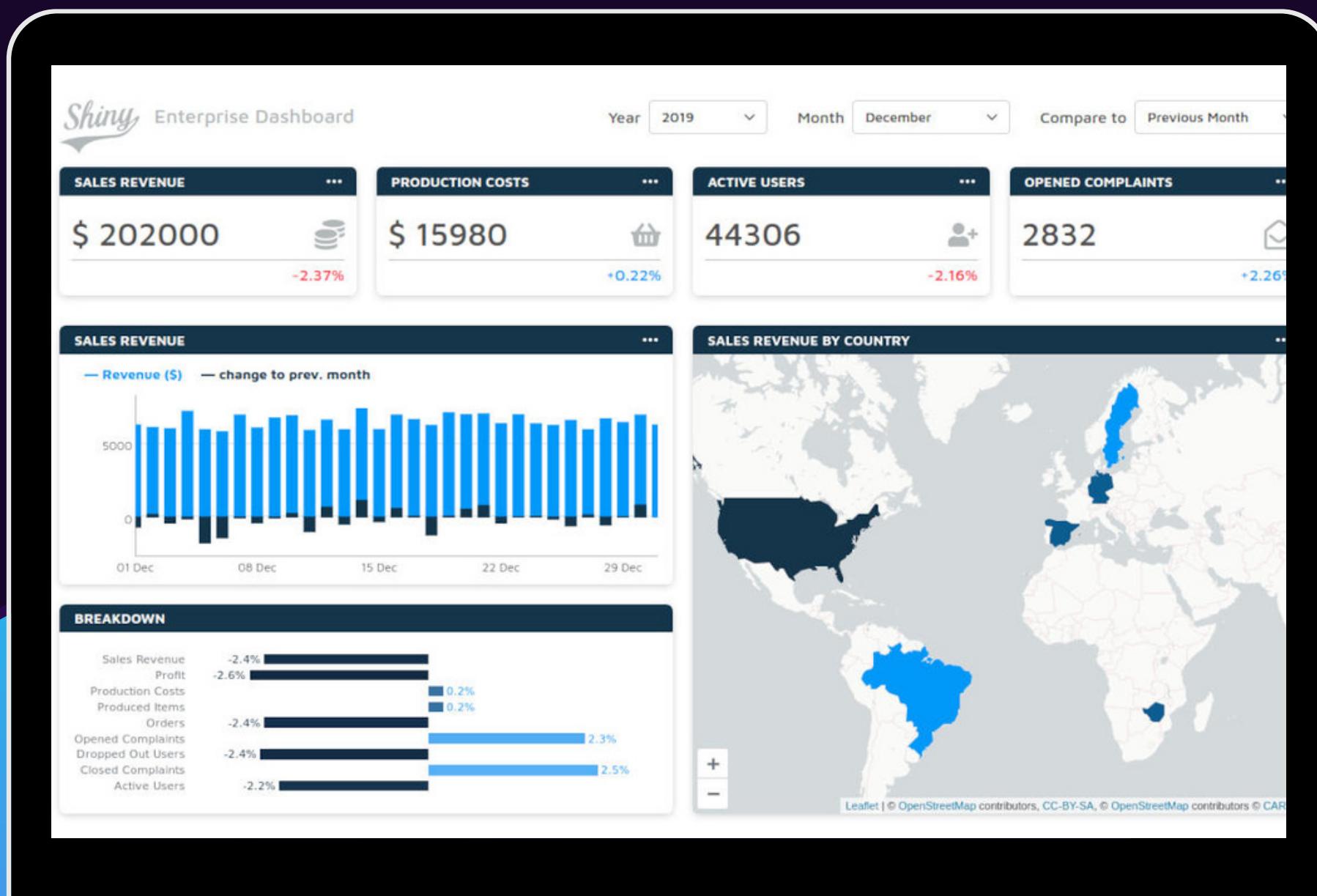
04

Rédaction d'un rapport d'analyse

Ce rapport doit permettre de présenter les données, les chiffres clés et les analyses effectuées. Une description et interprétation du modèle. Il est demandé de le rédiger au format Markdown. 5 à 8 pages sont attendues.



Les technologies et outils à utiliser



- Dash / Shiny
- Python
- Jupyter
- VS Code
- GitHub
- Draw.io



DEADLINE

Pour réaliser ce projet, votre équipe a :

- 7 x 3H

Vendredi 20 Octobre à 20H



Les étapes



Exploration, traitement et nettoyage des données



Création d'un repos GitHub et d'un environnement virtuel Python



Cartographie et extraction des données GPS via des API.



Modélisation du prix de vente et gestion des valeurs manquantes



Enrichissement avec des données OpenData pour améliorer l'apprentissage du modèle



Développement du squelette de l'application Dash



Développement des composants et test de déploiement.



Finalisation du modèle et développement de la page "prédiction" de l'application



Implémentation du module de réentrainement dans l'application et le développement de l'API



LIENS UTILES

- Cours Kaggle : <https://www.kaggle.com/learn>
- Crédation de l'environnement virtuel : <https://python.land/virtual-environments/virtualenv>
- Crédation d'une API : <https://towardsdatascience.com/deploy-a-machine-learning-model-using-flask-da580f84e60c>
- Déployer l'application : <https://render.com/> OU <https://shiny.posit.co/py/docs/deploy.html>
- Dessiner des schémas : <https://www.drawio.com/>
- Apprendre Shiny : <https://shiny.posit.co/py/>
- Cours de Python : <https://asardell.github.io/programmation-python/>
- Exemple de graphique en Python : <https://python-graph-gallery.com/>
- GitHub : https://github.com/asardell/stid1_2023/blob/main/TD1/td1.pdf

THANK YOU

SÉANCE 1

MATIN

- Installation de VS Code ,Jupyter et Python
- Compréhension du contexte
- Importation d'un jeu de données
- Statistiques descriptives sur
 - le nombre de ventes
 - évolution des ventes
 - top et flop des communes
 - les caractéristiques des ventes
 - surfaces
 - nombre de pièces
 - type local
- Compiler les 4 années d'historiques

APRÈS-MIDI

- Présentation de GitHub
- Crédit d'un compte GitHub
- Crédit d'un repository
- Crédit d'un environnement virtuel
- Représentation graphique avec et sans plotly
 - Nuage de points
 - Histogramme
 - Diagramme
 - Boxplot
- Cartographie avec le fichier commune

SÉANCE 2

MATIN

- Introduction Machine Learning
- Les métriques de classification
- Classification du type local pour réduire les valeurs manquantes
 - Arbre de décision
 - Régression logistique
 - KNN
 - Random Forest
- Interprétation du modèle avec SHAP

APRÈS-MIDI

- Les métriques de régression
- Régression sur le montant des valeurs foncières
- Régression linéaire
- Ridge, Lasso, ElasticNet
- Optimisation des modèles avec
 - Grid Search
 - Validation croisés
 - SMOTE
- Enrichir les modèles avec d'autres données (équipements communes, chômage, etc)