Sujet :

Détail:

TP Noté : À partir d'un jeu de données fourni vous devrez par l'intermédiaire de Scikit Learn ou de Dataiku réaliser la meilleure prédiction possible.
L'utilisation et la comparaison d'au moins 3 différents modèles de Machine Learning (régressions, modèle ensembliste) est attendue. Vous devrez en amont appliquer plusieurs étapes de data préparation (data cleaning, feature engineering) et de data visualisation.
Vous devrez mettre en place toutes les techniques d'entraînement vues en cours (Cross Validation, Random Search).
Vous utiliserez ensuite vos modèles pour faire des soumissions dans la compétition Kaggle associée au jeu fourni.

Equipe:

Vous pouvez travailler seul ou en équipe (2 maximum). Vous pouvez soit constituer une équipe mélangée de Dataiku et Python, soit rester sur une des deux solutions.

Contraintes:

Vous disposez de 2 jours de temps réservé pour réaliser ce projet.

Rendu:

Si vous choisissez de travailler directement en python nous attendons un document de type notebook commenté et illustré (graphique seaborn).
Pour les utilisateurs de Dataiku vous devrez fournir un document PDF expliquant votre démarche, agrémenté de screenshots du travail réalisé sur Dataiku (graphique, explication de vos étapes de data preparation, etc.)
Aussi, quelque soit les solutions que vous utiliserez un screenshot de votre note kaggle est demandé.

Enfin vous expliquerez votre démarche en séance lors d'une présentation de maximum 10 mn le vendredi 30 novembre..

Sujet proposé :

Catégorisez les crimes de San Francisco :
https://www.kaggle.com/c/sf-crime