CAHIER DES CHARGES

Projet de déploiement et de gestion du cycle de vie d'un modèle de Machine Learning

Fabrice BAZIN & Jonathan NAPOL

Étudiants chez DataScientest

Formation Machine Learning Engineer (MLOps)

Promotion de décembre 2022

DataScientest

TABLE DES MATIÈRES

Choix du sujet et du modèle

- Sujet : Prédire la gravité des accidents routiers en France.
- Modèle : Le modèle choisi est "XGBoost".

Définition des métriques et exigences de performances

- Métrique de performance : "f1-score"
- Exigences de performances : Un "f1-score" minimum de 0,8 et 0,5 respectivement pour les accidents de faibles et fortes gravité avec une "accuracy" d'au moins 0,8.

Schéma d'implémentation

- Collecte et nettoyage des données : Récupérer les données historiques sur les accidents routiers en France, nettoyer les données en éliminant les valeurs manquantes, les doublons, les valeurs aberrantes et les données inutiles.
- Extraction des caractéristiques : Sélectionner les variables pertinentes pour prédire la gravité des accidents, comme la localisation, la météo, le type de route, le type de véhicule impliqué, l'âge et le sexe du conducteur, etc.
- Préparation des données : Transformer les variables catégorielles en variables numériques et normaliser les données si nécessaire.
- Entraînement du modèle : Diviser les données en un ensemble d'entraînement et de test puis entraîner le modèle "XGBoost" sur l'ensemble d'entraînement en ajustant les hyperparamètres pour maximiser le "f1-score".
- Évaluation du modèle : Évaluer les performances du modèle sur l'ensemble de test en utilisant le "f1-score", ajuster les hyperparamètres si nécessaire.
- Scoring des zones à risque : Utiliser le modèle entraîné pour prédire la gravité des accidents dans différentes zones géographiques en fonction des informations météorologiques et géographiques.
- Comparaison avec les données historiques : Comparer les prédictions du modèle avec les données historiques pour évaluer la précision du modèle.

Récupération de nouvelles données

- Collecte et nettoyage des données : Récupérer les nouvelles données et les nettoyer en utilisant le même processus que pour les données historiques.
- Intégration des nouvelles données : Utiliser le modèle entraîné pour prédire la gravité des accidents en fonction des nouvelles données.
- Surveillance de la performance : Réévaluer les performances du modèle pour s'assurer que les exigences sont toujours satisfaites.