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 de la méthode

• Sujet : Prédire la gravité des accidents routiers en France.

• Méthode : Utiliser un modèle d'apprentissage automatique supervisé.

Définition des métriques et exigences de performances

• Métrique de performance : "accuracy", "precision", "recall" et "f1-score".

• Exigences de performances : Une "accuracy" globale d'au moins 90%, une "precision", un "recall" et un "f1-score" d'au moins 80% pour chaque catégorie de gravité d'accident.

Schéma d'implémentation

1. 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.
2. Exploration des données : Analyser les données pour identifier les tendances, les modèles et les relations entre les variables, et visualiser ces informations à l'aide de graphiques et de tableaux.
3. 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.
4. Préparation des données : Transformer les variables catégorielles en variables numériques, normaliser les données si nécessaire et gérer les données déséquilibrées, si pertinent.
5. Sélection du modèle : Comparer plusieurs modèles d'apprentissage automatique supervisé pour choisir celui qui offre les meilleures performances en fonction des métriques définies.
6. Entraînement du modèle : Diviser les données en un ensemble d'entraînement et de test puis entraîner le modèle choisi sur l'ensemble d'entraînement en ajustant les hyperparamètres pour atteindre les exigences de performances.
7. Évaluation du modèle : Évaluer les performances du modèle sur l'ensemble de test en utilisant l'"accuracy", la "precision", le "recall" et le "f1-score".
8. Validation croisée : Utiliser la validation croisée pour vérifier la robustesse et la généralisation du modèle sélectionné.
9. Scoring des zones à risque : Utiliser le modèle d'apprentissage automatique supervisé pour prédire la gravité des accidents dans différentes zones géographiques.
10. Comparaison avec les données historiques : Comparer les prédictions avec les données historiques pour évaluer la précision du modèle d'apprentissage automatique supervisé.

Récupération de nouvelles données et maintenance du modèle

1. Collecte et nettoyage des données : Récupérer les nouvelles données et les nettoyer en utilisant le même processus que pour les données historiques.
2. Intégration des nouvelles données : Utiliser le modèle d'apprentissage automatique supervisé entraînée pour prédire la gravité des accidents en fonction des nouvelles données.
3. Surveillance de la performance : Réévaluer les performances du modèle d'apprentissage automatique supervisé pour s'assurer que les exigences en matière d'"accuracy", de "precision", de "recall" et de "f1-score"sont toujours satisfaites.
4. Réentraînement et ajustement du modèle : Si les performances du modèle se détériorent ou si de nouvelles données pertinentes deviennent disponibles, réentraîner le modèle en intégrant les nouvelles données et en ajustant les hyperparamètres, si nécessaire, pour maintenir ou améliorer les performances. Effectuer cette étape régulièrement pour assurer la précision et la pertinence du modèle au fil du temps.
5. Documentation et communication : Documenter toutes les étapes du processus, y compris la sélection du modèle, les choix d'hyperparamètres, les performances obtenues et les ajustements apportés au modèle. Communiquer ces informations aux parties prenantes concernées pour assurer la transparence et faciliter la collaboration.
6. Plan de déploiement : Définir un plan pour déployer le modèle d'apprentissage automatique supervisé dans un environnement de production, y compris l'intégration avec les systèmes existants, les interfaces utilisateur et les processus métier.
7. Formation des utilisateurs : Former les utilisateurs finaux et les parties prenantes sur l'utilisation du modèle, les résultats qu'il produit et les limites potentielles de ses prédictions.
8. Plan de suivi et d'amélioration continue : Établir un plan pour surveiller les performances du modèle, recueillir les commentaires des utilisateurs et des parties prenantes, et apporter des améliorations continues au modèle et au processus de prédiction de la gravité des accidents routiers en fonction des besoins identifiés.