

Guide de Réalisation : Projet 7 Data Scientist d'OpenClassrooms

1. Comprendre le Contexte et les Données

- Objectif : Créer un modèle de scoring de crédit qui aide à prendre des décisions d'octroi de crédit. - Données : Généralement, les données de ce projet proviennent d'une base de données de clients, avec des informations sociodémographiques, financières et des historiques de crédits. - Exploration des données : Analysez chaque variable pour comprendre ses caractéristiques, son importance, et nettoyez les données en cas de valeurs manquantes ou incohérentes.

2. Prétraitement des Données

- Nettoyage : Imputez les valeurs manquantes et traitez les données aberrantes. - Encodage : Transformez les variables catégorielles (si présentes) en valeurs numériques (one-hot encoding ou encodage ordinal). - Normalisation / Standardisation : Standardisez ou normalisez les variables numériques pour faciliter l'apprentissage du modèle.

3. Analyse Exploratoire des Données (EDA)

- Visualisation des distributions : Étudiez la répartition des caractéristiques (histogrammes, pair plots). - Analyse des corrélations : Identifiez les relations entre les variables pour mieux comprendre les interactions (matrice de corrélation). - Feature Engineering : Créez de nouvelles variables, si pertinent, en combinant les variables existantes ou en appliquant des transformations.

4. Sélection des Caractéristiques

- Sélection de variables : Utilisez des méthodes de sélection de caractéristiques (comme RFE ou l'analyse de l'importance des variables) pour réduire la dimensionnalité. - Test d'importance : Entraînez un modèle simple (ex : Random Forest) pour vérifier l'importance de chaque variable.

5. Choix et Entraînement du Modèle

- Sélection du modèle : Pour ce type de projet, des algorithmes comme la régression logistique, les forêts aléatoires (Random Forest) ou les modèles basés sur les arbres (XGBoost, LightGBM) sont souvent adaptés. - Entraînement du modèle : Entraînez votre modèle sur les données prétraitées, en ajustant les hyperparamètres. - Validation croisée : Utilisez la validation croisée pour évaluer les performances et éviter le surapprentissage.

6. Évaluation du Modèle

- Métriques de performance : Évaluez le modèle avec des métriques de classification (ex : AUC, précision, rappel, F1-score). - Courbe ROC et AUC : Visualisez la courbe ROC et calculez l'AUC pour mesurer la capacité discriminante du modèle. - Interprétabilité : Utilisez des méthodes comme SHAP ou LIME pour expliquer les prédictions et mieux comprendre l'impact de chaque variable.

7. Optimisation du Modèle

- Ajustement d'hyperparamètres : Utilisez des techniques comme Grid Search ou Randomized Search pour optimiser les hyperparamètres. - Ensemble de modèles : Si pertinent, combinez plusieurs modèles (stacking, bagging) pour améliorer la robustesse.

8. Création d'une Interface de Visualisation

- Tableau de bord interactif : Implémentez un tableau de bord (par ex., avec Streamlit, Dash ou Flask) pour visualiser les prédictions, les scores de risque, et permettre aux utilisateurs de tester avec des données de nouveaux clients. - Explicabilité : Intégrez les graphiques d'explicabilité pour chaque prédiction (graphiques SHAP, LIME) et permettez à l'utilisateur de comprendre pourquoi un client reçoit un certain score.

9. Documentation et Rapport

- Rapport final : Documentez chaque étape du projet dans un rapport clair et structuré. Expliquez les choix d'algorithmes, les étapes de prétraitement et d'analyse, et les résultats. - Précisez les limites : Indiquez les limites du modèle et les axes possibles d'amélioration.

Conclusion

En suivant ces étapes, vous pourrez réussir le projet de scoring de crédit et proposer une solution complète et interprétable pour une institution financière.