Guide d'étapes clés : Construisez un modèle de scoring

Comment utiliser ce document ?

Ce guide vous propose un découpage du projet en étapes. Vous pouvez suivre ces étapes selon vos besoins. Dans chacune, vous trouverez :

- des recommandations pour compléter la mission ;
- les points de vigilance à garder en tête ;
- une estimation de votre avancement sur l'ensemble du projet (attention, celui-ci peut varier d'un apprenant à l'autre).

Suivre ce guide vous permettra ainsi :

- d'organiser votre temps ;
- de gagner en autonomie ;
- d'utiliser les cours et ressources de façon efficace ;
- de mobiliser une méthodologie professionnelle que vous pourrez réutiliser.

Gardez en tête que votre progression sur les étapes n'est qu'une estimation, et sera différente selon votre vitesse de progression.

Recommandations générales

- Assurez-vous de comprendre la problématique métier et l'objectif du projet avant de commencer.
- Familiarisez-vous avec les méthodologies et techniques utilisées dans l'apprentissage supervisé en Machine Learning.
- Veillez à acquérir une compréhension technique des différents types de modèles de Machine Learning couramment utilisés pour la tâche identifiée.
- Familiarisez-vous avec le jeu de données et les variables disponibles pour la modélisation.

- Gardez à l'esprit l'importance de l'interprétabilité du modèle pour les utilisateurs finaux.
- Soyez prêt à justifier vos choix tout au long du projet, en tenant compte du contexte métier.

Étape 1 : Choisissez un kernel et effectuez un analyse exploratoire

10% de progression

Avant de démarrer cette étape, je dois avoir :

- Compris les exigences du projet et les objectifs du modèle de scoring.
- Revu les ressources fournies pour vous aider à choisir un kernel Kaggle pertinent.

Une fois cette étape terminée, je devrais avoir :

- Effectué une analyse exploratoire du jeu de données pour comprendre sa structure et ses caractéristiques.
- Identifié des opportunités de feature engineering pour améliorer la performance du modèle.

Recommandations :

- Utilisez le kernel recommandé pour la préparation des données et l'analyse exploratoire.
- Créez au moins trois nouvelles variables pertinentes à partir des variables existantes.
- N'hésitez pas à adapter le kernel en ajoutant des commentaires pour expliquer vos actions.

Points de vigilance :

- Assurez-vous d'utiliser le bon fichier (application_train.csv) pour l'analyse exploratoire.
- Vérifiez les valeurs manquantes et les valeurs extrêmes dans le jeu de données.

Ressources :

- Appuyez-vous sur le cours Nettoyer et analysez votre jeu de données pour la méthodologie d'analyse.

- Utilisez le kernel start-here-a-gentle-introduction pour l'exploration et l'inspiration de techniques de feature engineering.

Étape 2 : Définissez une métrique adaptée à la problématique métier

30% de progression

Avant de démarrer cette étape, je dois avoir :

- Compris l'importance du coût des erreurs de prédiction dans le contexte financier.
- Compris comment créer une fonction de coût métier pour évaluer les performances des modèles.

Une fois cette étape terminée, je devrais avoir :

- Mis en place une fonction de coût métier prenant en compte les coûts des faux positifs et des faux négatifs.
- Calculé un score "métier" pour évaluer les performances des différents modèles.

Recommandations :

- Utilisez la fonction de coût métier pour optimiser les hyper paramètres des modèles.
- Assurez-vous de calculer le seuil optimal pour le score "métier".

Points de vigilance :

- Assurez-vous de comprendre le coût des erreurs de prédiction dans le contexte financier.
- Vérifiez que votre score "métier" est calculé correctement en prenant en compte les faux positifs et les faux négatifs.

Ressources :

- Appuyez-vous sur le cours evaluez les performances d'un modèle pour comprendre comment créer un score "métier" adapté.

Étape 3 : Optimisation et évaluation des modèles de Machine Learning

40% de progression

Avant de démarrer cette étape, je dois avoir :

- Compris l'importance de tester différents types de modèles pour le problème identifié.
- Etudié les différentes ressources théoriques mises à disposition.

Une fois cette étape terminée, je devrais avoir :

- Testé et comparé différents modèles, des plus simples aux plus complexes.
- Utilisé des techniques pour gérer le déséquilibre des classes. Maîtrisé l'utilisation des pipelines scikit learn pour l'optimisation et l'évaluation des modèles

Recommandations :

- Utilisez la Cross-Validation pour évaluer les performances des modèles de manière robuste.
- Explorez différentes valeurs d'hyper paramètres pour obtenir des modèles performants à l'aide de Grid Search.
- Synthétiser vos résultats (tableau comparatif + courbe ROC)

Points de vigilance :

- Assurez-vous de tester une variété de modèles, des modèles simples aux modèles plus complexes.
- Vous devrez veiller à équilibrer vos classes. Vous devrez justifier votre démarche.
- Des scores AUC supérieurs à 0.82 pourraient révéler une fuite de données dans le pipeline mis en place.

Ressources :

- Utilisez les ressources externes et les cours recommandés pour comprendre et réaliser le benchmark des modèles
- Utilisez la fonction pipeline fournie par scikit learn
- Rééquilibre vos classes grâce à la librairie Imbalanced-learn

Étape 4 : Développez un module d’explicabilité de votre modèle de Machine Learning

80% de progression

Avant de démarrer cette étape, je dois avoir :

- Compris l'importance de l'interprétation du modèle pour les utilisateurs finaux.
- Revu les concepts de feature importance globale et locale.

Une fois cette étape terminée, je devrais avoir :

- Utilisé des librairies spécialisées pour calculer la feature importance globale et locale.
- Fourni des informations pour expliquer les prédictions du modèle aux chargés d'étude.

Recommandations :

- Assurez-vous de comprendre comment interpréter les résultats de l'analyse.

Points de vigilance :

- Soyez prêt à expliquer aux chargés d'étude comment les features influencent les prédictions du modèle.

Ressources :

- Utilisez les librairies SHAP ou LIME pour comprendre et mettre en œuvre l'analyse de la feature importance.

Projet terminé !