Test Technique Data Science

El Hadji Gagny

Sommaire

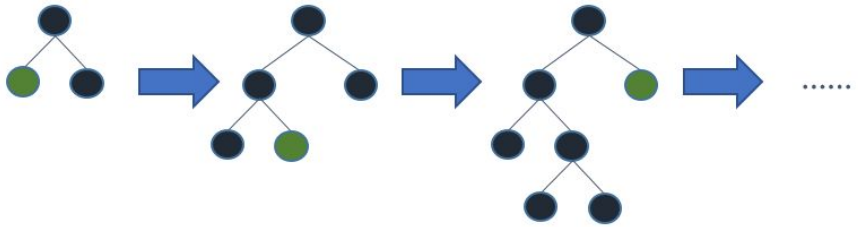
Objectifs globaux

- Extraction et traitement de la donnée
- Data visualisation et tests statistiques
- Techniques de transformation pour la création de variables métiers
- Construction du modèle de prédiction (classification binaire)
- Interprétabilité du modèle

Données

- Tables de près de 200 000 instances de candidatures sur 11 variables (age, salaire, nombre d'années d'expérience, ...)
- 11 % des candidatures acceptées (déséquilibre des classes)
- Variables faiblement dépendantes et peu corrélées
- l'âge, l'expérience , le salaire avec des distribution superposables suivant la population (candidature acceptée, ou refusée)
- Variable couleur des cheveux non retenu dans le modèle
- Création de nouvelles variables de profilage en fonction de l'âge et de l'expérience
- Création de variables de temporalité (mois, jour de la semaine, ...) pour prendre compte de possibles répercussions d' évènement boursiers (cours de l'or) sur le recrutement
- Suppression de lignes corrompues (âge inférieur à expérience)
- Traitement et suppression de valeurs aberrantes
- Label encoding des variables (s'adapte bien avec le lightgbm)
- Valeurs manquantes peu présentes (un maximum de 9% sur certaines colonnes), remplacement par la moyenne.

Modèle (LightGBM)



Leaf-wise tree growth

- **Rapidité du training** : Grâce à la profondeur Leaf-wise et support du parallélisme
- **Efficacité** : Moindre utilisation de mémoire
- **Explicabilité** : Grâce au support et splitting du dataset volumineux à chaque noeud
- **Meilleure accuracy** par rapport aux autres algorithmes de boosting
- **Performant** avec de larges datasets
- Précision : **Accuracy** : 0.81

Résultats

Résultats sur ensemble de validation:

- accuracy: 82 %
- roc : 75 %

Résultats sur jeu de test:

- accuracy: 81%
- roc : 63%

- Modèle peu sujet à du sur-apprentissage (early-stopping)
- Rééchantillonnage des données
- Performance stable
- Le salaire, la note et l'âge semble très importantes sur l’issue de la prédiction

Limites et Extensions

Limites

- **Présence de variables de biais (profilage en fonction de l'âge)**

Extensions

- **Usage de méthodes de sampling géométriques (SMOTE)**
- **Batch Optimization à la place du Random Search pour une exploration plus poussée des paramètres**
- **Pénaliser l’erreur avec sample weight (Lightgbm parameter)**
- **Aggregation avec des tables externes (données sur le cours de l’or pour rechercher des corrélations avec la date du recrutement)**
- **Métrique basée sur les erreurs de type 1 et 2 (Precision, Recall)**