Test technique Quantmetry - OrFée

Alban de Crevoisier

Abstract

Il s'agit ici de prédire le succès ou l'échec d'une candidature au poste de chercheur d'or chez OrFée.

1 Introduction

Ce test a été réalisé en python 3.7.3 avec les bibliothèques pandas pour le data processing, scipy et scikit-learn pour le machine learning, et matplotlib et seaborn pour la visualisation des résultats.

Tout le code est fourni en annexe, dans le fichier orfee.py accompagné d'un README.

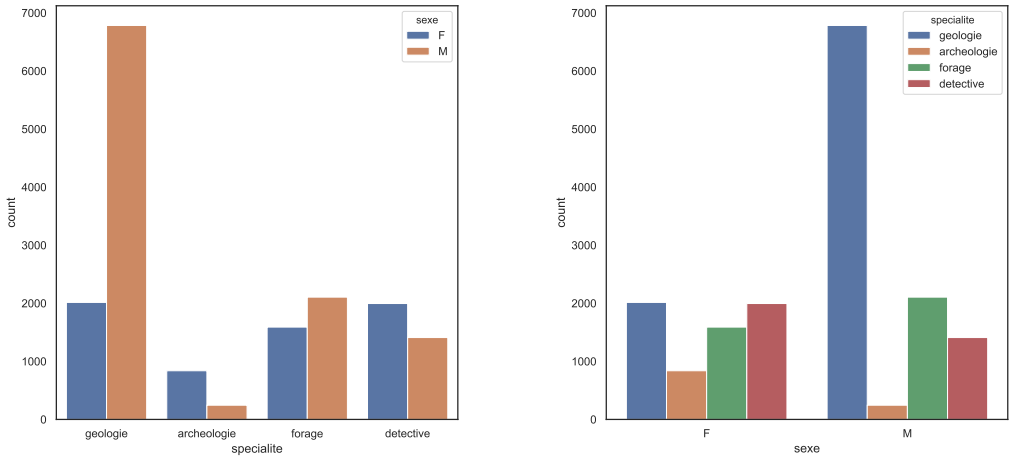
2 Statistiques Descriptives

2.1 Jeu de données

Le jeu de données comporte 20 000 observations, dont 16 984 sont parfaitement bonnes - pas d'élément manquant ni de valeur en dehors de leurs contraintes, comme un âge négatif par exemple. J'ai décidé pour ce test de ne conserver que les bonnes valeurs, estimant que la perte d'environs 15acceptable.

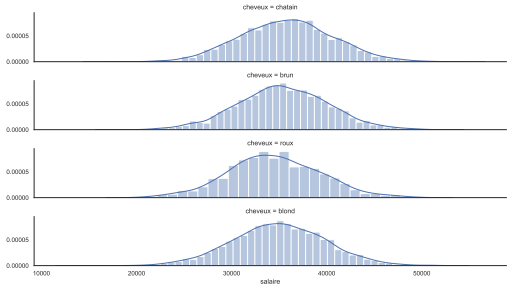
2.2 Dépendances Statistiquement Significatives

2.2.1 Spécialité et Sexe

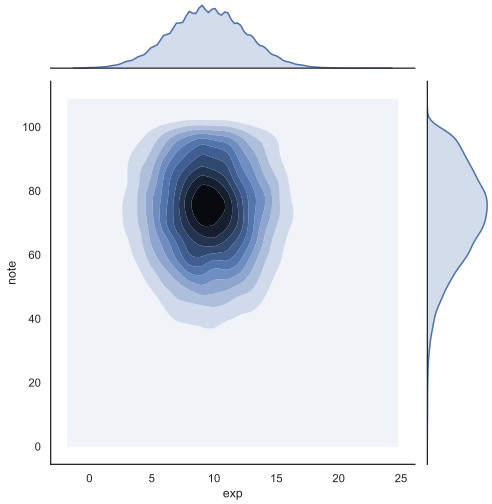


On observe bien sur ce graphique une corrélation assez forte entre spécialité et sexe. Elle est de 0.32 pour la géologie et 0.21 pour l'archéologie et le métier de détective.

2.2.2 Couleur de cheveux et Salaire demandé



2.2.3 Expérience et note

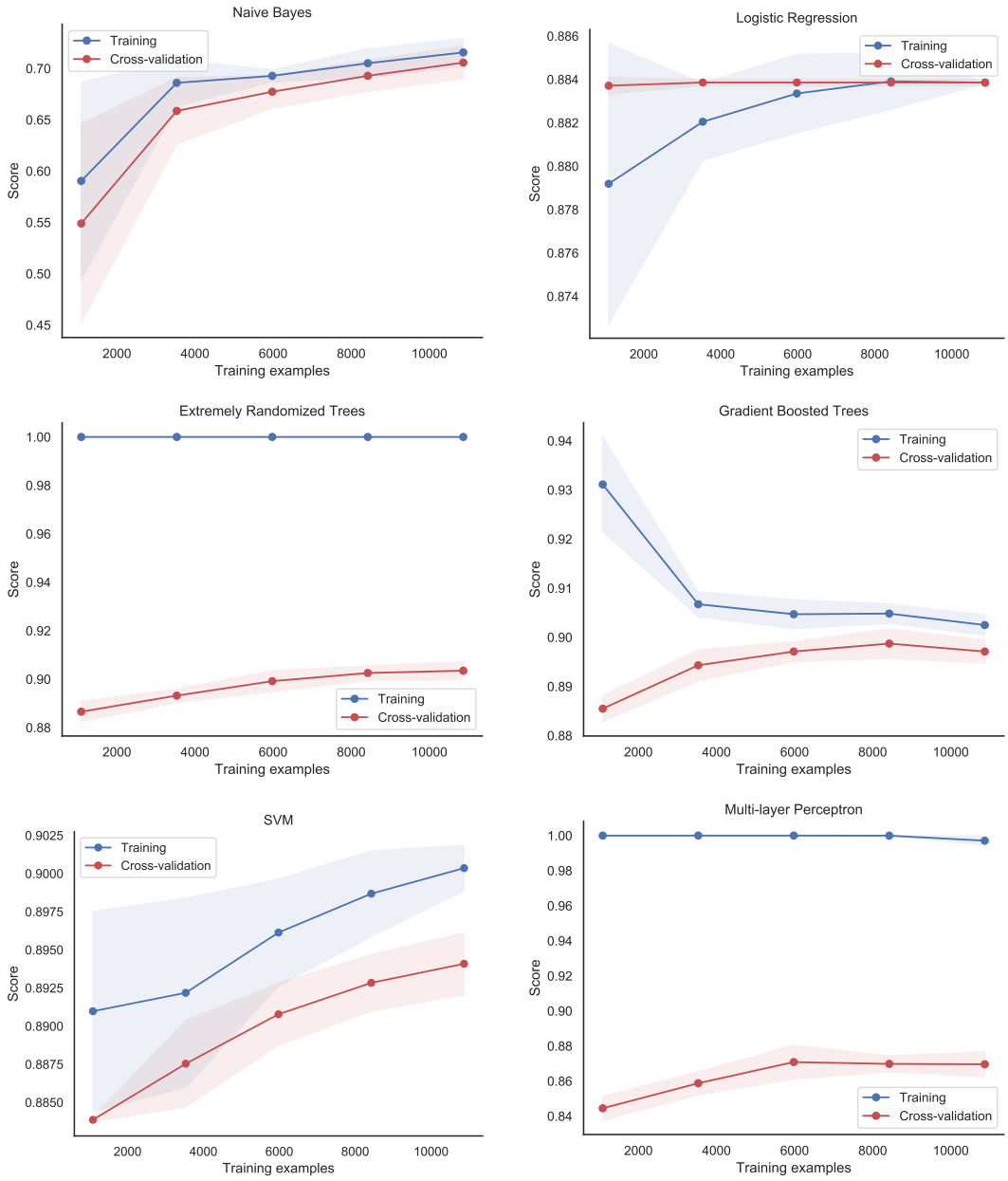


Corrélation: -0.016512926097244886

3 Machine Learning

3.1 Sélection de modèle

Commençons par comparer quelques modèles classiques avec les paramètres par défaut pour se faire une première idée : un classificateur bayesien standard,une régression logistique, une random forest, une SVM et un réseau de neurones dense.



Le classificateur bayesien fournit un seuil minimal de performances attendues, que tous les autres classificateurs dépassent, ce qui témoigne de leur efficacité.
La régression logistique ne semble pas très prometteuse, puisqu'elle a déjà l'air d'avoir convergé vers un résultat moyen.

La forêt d'arbres décisionnels overfit énormément mais fournit tout de même un bon résultat, en augmentant le biais il devrait être possible d'en améliorer les performances.
Les arbres avec gradient boosting ont au contraire l'air d'underfit, peut-être qu'améliorer la variance pourrait fournir de bons résultast.
La SVM continue manifestement à bénéficier des nouvelles données, essayer de nettoyer les données mises de côté ou d'accélerer l'apprentissage devrait fournir une nette amélioration. Cependant, elle ne converge actuellement pas vers une valeur très intéressante.
Le réseau de neurones overfit beaucoup et fournit une prédiction correcte, mais au prix de beaucoup plus de temps d'exécution
Au vu de ces résultats, essayons d'améliorer les performances de la forêt d'arbres de décision et des arbres avec gradient boosting.

3.2 Optimisation des hyperparamètres

foo