Bilan du projet Python pour un Data Scientist - Olivier Supplisson

Ce projet a été très enrichissant. D'un point de vue « humain », ce fut un véritable plaisir de travailler avec Nicolas. Nos deux profils étaient très complémentaires : mon domaine de compétence se situe plus dans le choix des algorithmes, l'analyse et la rédaction que dans l'implémentation pure. Contrairement à Nicolas qui code de manière très propre et efficace. Cette complémentarité a rendu le travail extrêmement fluide et agréable et m'a permis de beaucoup apprendre en terme de code, notamment la syntaxe des classes que je ne connaissais pas.

Ayant déjà eu l'occasion d'appréhender `ScikitLearn` et plus généralement les algorithmes de *machine learning* durant le projet et mon stage de première année, ce projet a été l'occasion d'approfondir ma compréhension pratiques de ces algorithmes. Il m'a également donné l'occasion d'approfondir le sujet de la sélection des *features* et m'a permis de découvrir l'algorithme `Boruta` qui se révèle très pratique, notamment car très facile d'implémentation et reposant sur des principes théoriques relativement intuitifs, pour faire un premier tri dans une base de données ayant beaucoup de variables.

De manière plus générale, ce projet m'a permis de me perfectionner dans la manipulation et la représentation de données dans `Python`. Il m'a également poussé à réfléchir davantage sur la manière de prendre en compte le déséquilibre d'effectif entre les différentes classes de l'*outcome*. Le fait que notre solution en deux étapes n'ait pas marché fut une déception car je la trouvais relativement intuitive et bien pensée.

Malgré toutes ces notes positives, la base que nous avons choisi ne nous a pas donné l'opportunité de nous confronter à la tâches extrêmement importante qu'est le *features engineering*. Bien sûr, nous avons ajouté les croisements au premier ordre, tout comme nous avons considéré les variables centrées réduites mais à aucun moment nous n'avons du nettoyer la base, celle-ci étant prête à utiliser. En particulier, nous n'avons eu aucun problème de traitement des valeurs manquantes, qui est, avec la sélection de variables, un des problèmes majeurs auquel fait face le *data scientist* et qui est très complexe à traiter.

Je reste également circonspect sur l'intérêt de notre travail, outre celui de nous permettre de manipuler des algorithmes et de coder. En effet, je trouve notre sujet (le vin) loin d'être original. Ainsi, pour mes projets de 3A, je ferai attention à choisir des sujets à la fois complexes du point de vue du *features engineering* et original de par le thème abordé. Malgré tout, ce projet aura été très formateur et je ne regrette pas notre choix. Je suis également très content de notre rendu et de la philosophie avec laquelle nous avons codé : notre code est réutilisable pour n'importe quelle base de données et permet donc d'obtenir des premiers résultats exploratoires très rapidement, y compris quand le nombre de variables est très important (grâce à Boruta). Toutes les classes de notre projet pourront donc être utilisées lors de travaux ultérieurs : à défaut d'être intéressant, notre travail aura donc le mérite d'être utile et de nous éviter de tout avoir à recoder ultérieurement !