SD701 : Prédiction de la « Draft » NBA

1. Description du projet

Aux États-Unis, pour qu'un athlète puisse devenir professionnel, il doit s'inscrire dans une université. Après la validation de la première année, il est éligible à passer professionnel. Pour faire ses preuves, il participe, avec son équipe d'université, à la compétition organisée par la NCAA (*National Collegiate Athletic Association*). Le sport choisi pour ce projet est le basketball et sa ligue, la NBA (*National Basketball Association*). Quand il souhaite faire le grand saut, il s'inscrit à la « draft ».

La draft est un évènement annuel où sont réunis le commissionnaire de la ligue et les dirigeants des trente équipes. Chaque équipe va sélectionner, selon un ordre précis, un joueur s'ayant inscrit à la draft. Au total, seulement 60 joueurs rejoignent la ligue chaque année sur environ 3500 inscrits.

Les 30 premiers journées draftés lors de la NBA Draft de 2019

L'objectif de ce projet est de déterminer si un universitaire peut potentiellement rejoindre la NBA suivant ses statistiques et ses performances dans la compétition NCAA.

Pour répondre, on réalisera une première phase de « web scrapping » des statistiques des joueurs. La deuxième phase consistera à l'analyse des statistiques et la troisième phase sera consacrée à trouver le meilleur modèle de classification de Machine Learning.

2. Collectage de données

2.1. Extraction des données

La collecte des données était assez complexe car il n'existe pas de datasets répondant à cette problématique. Il existe une compétition Kaggle organisée par la NCAA et Google Cloud mais le dataset, bien que fourni, donne des informations que sur les salles de sport, l'affluence et sur les matchs.

Le choix s'est porté sur la référence en statistiques sportives, Basketball Reference. Cependant, ce site ne met pas à disposition une API. L'extraction de la donnée a été faite par *Beautiful Soup* complété par une API trouvé sur Github.

2.2. Nettoyage des données

Un autre problème encouru est le mapping entre les joueurs NCAA et les joueurs draftés. En effet, la base de données des joueurs NBA a été faite des années avant celle des joueurs universitaires. De ce fait, les joueurs se trouvant à la fois dans la base de données NBA et NCAA ne possèdent pas le même « id ». Pour créer dans notre dataset la colonne « y » à prédire (« drafté » : 0 non, 1 oui), il a fallu faire une jointure entre les noms des joueurs qui a résulté un nouveau problème : les joueurs homonymes.

La présence de données dupliquées devait être traitée. Prenons le cas d'un joueur X qui décide de ne s'inscrire à la draft qu'au bout de sa quatrième année (soit à la fin du cursus universitaire). Le joueur X possèdera dans le dataset 8 lignes ! En effet, il aura les 4 lignes dédiées à ses performances sur ses 4 années puis il aura des lignes appelées « carrière » qui est la somme des statistiques après chaque année :

1ère ligne carrière : Somme des stats de la première année universitaire
2ème ligne carrière : Somme des stats de la 1ère et 2ème années universitaires
3ème ligne carrière : Somme des stats de la 1ère, 2ème et 3ème année universitaire
4ème ligne carrière : Somme des stats de la 1ère, 2ème, 3èmes et 4èmes années universitaires

Pour la suite du projet, un joueur universitaire sera assigné à ses statistiques en carrière de sa dernière année universitaire.

Dernier problème, certaines métriques que l'on appelle « stats avancées » ne sont pas présentes pour tous les joueurs et notamment les joueurs les plus méconnus. L'une des stratégies serait de remplacer le NaN par la moyenne mais pour obtenir une meilleure analyse, il est possible de calculer ces métriques à l'aide de d'autres statistiques.

Exemple du « true shooting pourcentage » ou « pourcentage au tir réel » qui représente l'efficacité réelle d'un tireur. Cette métrique peut être utilisé à l'aide de PTS (moyenne de points marquées par match), FGA (moyenne de tirs tentées par match) et le FTA (moyenne de lancer francs tentées par match)

$$TS\% = \frac{PTS}{2(FGA+(0.44\times FTA))} \times 100\%$$

3. Exploitation des données

3.1 Analyse des données

Le dataset des joueurs universitaires entre 2010 et 2019 montre que le processus de draft est très sélectif : *seulement 2.3% des universitaires deviennent professionnels*.

L'objectif cette analyse est de tenter de trouver les critères et le pattern de sélection :

Conférence : la première division universitaire est composée de 32 conférences (qui sont composées chacune d'universités). Ces conférences ont une répartition homogène de joueurs (600 en moyenne). Cependant, 5 conférences possèdent presque 75% devenus professionnels

Equipes : Parmi les 5 conférences cités, 4 universités se partagent 26% des futurs professionnels

Postes : il existe 5 différentes positions au basketball, en universitaire on en décompte 3. On remarque que la répartition devenus professionnels est différente de celle de tous les joueurs.

Les joueurs occupant le rôle de « Forward » et « Center » ont une plus grande importance dans le sous échantillon comparé à « Guard ». Ces rôles ont tendance à être attribués aux joueurs très grand de tailles

Taille : Les courbes de densité de fréquence montrent que les joueurs les plus grand de tailles sont privilégiés dans la sélection à la draft

Statistiques de bases : Il existe 5 statistiques de base au basketball : points, passes décisives, contres, rebonds et interceptions

En réalisant des courbes de densité pour chaque statistique, on se rend compte que seulement 2 statistiques peuvent être des critères de sélection (points et rebonds).

Statistiques avancées : Des métriques inventés par des statisticiens et qui permettent d'évaluer une performance d'un joueur sur son impact offensif ou défensif, son efficacité au tir … Ces métriques sont de donc de bons critères de sélection

Classement TOP 100 à la sortie du lycée : Un critère très intéressant car les 50.4% des 100 meilleurs universitaires à la sortie du lycée deviennent professionnels.

Ainsi cette étude réalisée sur les statistiques des joueurs universitaires permet de conclure qu'il est possible de déterminer si un universitaire pourrait devenir professionnel s'il est plus grand que la moyenne à son poste, joue dans l'une des 5 meilleurs conférences de première division dans une université réputée tout en étant un très bon joueur à sa sortie de lycée.

3.2 Prédictions de la sélection de draft

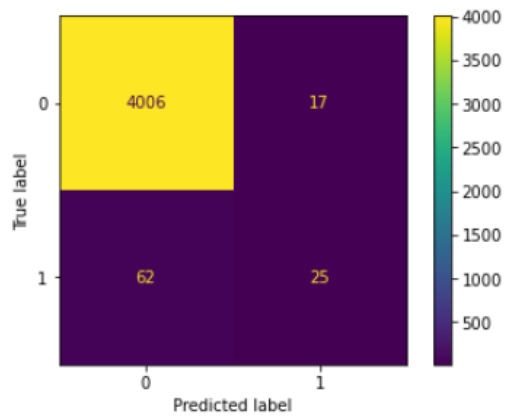
L'objectif est de trouver le meilleur modèle de classification binaire possible de Machine Learning pour prédire la sélection à la draft d'un joueur universitaire.

Le dataset est composé d'environ 20 000 joueurs de la première division de la NCAA entre 2010-2019.

Les labels à prédire sont : 1 le joueur a été drafté, 0 le joueur n'a pas été drafté

Le premier modèle utilisé est la régression logistique avec des hyperparamètres par défaut. L'accuracy obtenue sur ce modèle est très bon (0.98). Bien qu'on remarque une accuracy à 0.98, le recall et le f1-score sont très faible (resp. 0.29 et 0.39).

La métrique à ne pas prendre en considération est l'accuracy car nous sommes en situation d'unbalanced dataset (98% de non draftés pour 2% de draftés). Dans notre cas, le recall est plus intéressant. Pour bien estimer la performance d'un modèle, on devrait s'intéresser au F1-score qui est un bon compromis entre le recall et la précision.



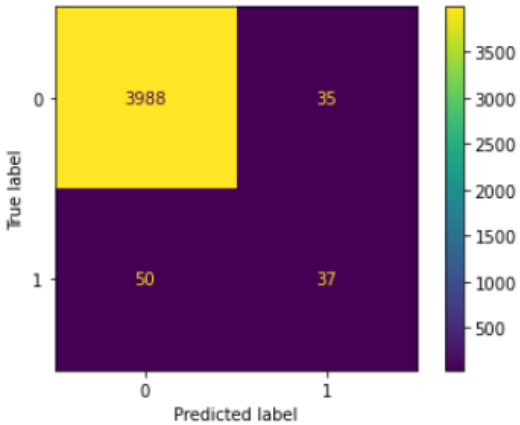
Modèle 1

Matrice de confusion pour la régression logistique de base, ce modèle n'est pas performant car le nombre de faux négatif est beaucoup plus élevé que de vrais positifs.

GridSearch_CV

Pour tenter de trouver le meilleur modèle possible, on va utiliser un GridSearch_CV en utilisant un cross validation à 5 folds et en utilisant 4 algorithmes : Régression logistique, KNN, Arbre de décisions et SVM avec des paramètres différents.

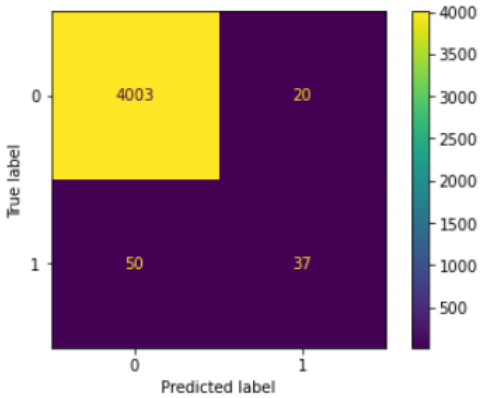
Ce Gridsearch nous donne comme meilleur modèle un SVC avec comme paramètres : {'C': 100, 'gamma': 0.0001, 'kernel': 'rbf'} et avec un score f1 de 0.49



Modèle 2 : Matrice de confusion après Grid Search CV, on observe une augmentation de vrais positifs et une baisse de faux négatifs, ce qui est encourageant. Cependant, le nombre de faux positifs augmente.

Tentons de trouver un meilleur modèle.

Comme première tentative d'amélioration du modèle, les features ont été standardisé puis on a réalisé un ACP. Enfin les features représentant la conférence et l'équipe du joueur qui étaient encodés à l'aide de *LabelEncoding* on était encodé, cette fois ci, par *One Hot Encoder*.

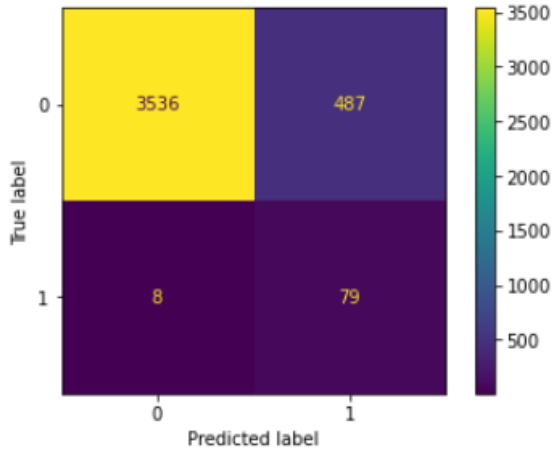


Modèle 3 : Le nombre de faux positifs baisse. Cependant, le nombre de faux négatifs reste contant. C'est une légère amélioration comparée au modèlé précédent.

Pour la dernière tentative d'amélioration, utilisons des algorithmes d'*undersampling* qui sont utilisés dans les *unbalanced datasets*

Le principe est de créer un sous dataset ayant une proportion de 50% le label 1 et 50% le label 0 (contre le 98%-2% du dataset de départ). Par la suite, on réalise un gridsearch pour trouver le meilleur modèle. Ce modèle qui est entrainé sur le sous echantillon sera utilisé pour prédire le dataset de test de l'échantillon de base.

Les deux meilleurs modèles sont la régression logistique et un arbre de décision. Utilisons le modèle de Logistic Regression avec comme paramètres {'C': 1000.0, 'penalty': 'l2'} et un f1-score de 0.92



Modèle 4 : On remarque que le taux de faux négatifs a considérablement chuté tandis que celui des vrais positifs à énormément augmenté. C'est ce qui était recherché.

Cependant, un inconvénient est l'explosion du nombre de faux positifs.

Conclusion

Le choix du meilleur modèle entre les deux derniers doit être une décision du Data Scientist et du use case. En effet, si la situation était le dépistage d'une maladie, alors le dernier modèle est le plus intéressant car il réduit considérablement les faux négatifs. Le nombre de faux positifs ne serait pas problématique.

Si, par contre, l'use case était la détection de spam dans une boite de messagerie, le dernier modèle va considérer beaucoup de mails comme spam donc certains emails importants ne seront pas consultés par l'utilisateur.

Dans le cas de la prédiction de la draft, le quatrième modèle n'est pas intéressant car il donne 487 faux positifs.

Le nombre de cas de faux négatifs du troisième modèle pourrait être expliqué par des données non présentes dans le dataset comme l'état d'esprit du joueur, la forme de l'équipe, la présence de meilleurs joueurs dans l'équipe universitaire, des blessures, les distinctions personnelles, un potentiel caché du joueur, le besoin des équipes professionnels, les jugements biaisés des recruteurs etc

Bibliographie :

Scrapping :

Documentation de l'API : https://sportsreference.readthedocs.io/en/stable/ncaab.html#player

Statistiques des joueurs NBA et NCAA : https://www.basketball-reference.com/about/ratings.html

Exploration de données

https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

Machine Learning:

Explication des métriques de la classification : https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

Algorithmes pour Unbalanced Datasets : https://www.kaggle.com/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets

Glossaire

NBA	Fédération professionnelle de basketball américaine
NCAA	Fédération de basketball universitaire
Draft	Processus de sélection des joueurs universitaires dans le but de devenir professionnel
Rookie	Joueur professionnel ayant moins d'une année d'expérience professionnelle