Projet Deep Learning
Prédiction du nombre de passagers

LAVAL Léa - GUEDIDI Feriel

10/02/2021

I. Introduction

Le but du projet est de construire un modèle qui permette de prédire avec le plus de précision possible le nombre de passagers à bord d'avions qui décollent et atterrissent de vingt différents aéroports américains. Pour cela nous avons à disposition une base de données d'entraînement contenant initialement les variables suivantes :

- Les aéroports de départ et d'arrivée : Atlanta, Los Angeles, Chicago, Dallas, Denver, New York, San Francisco, Seattle, Las Vegas, Orlando, Charlotte-Douglas, Phoenix, Miami, Houston, Philadelphie, Detroit, Minneapolis, New York-Laguardia, Boston.
- La date de départ (année-mois-jour)
- Le nombre de semaines moyen entre le jour de réservation et le jour de départ
- L'écart type de la variable précédente, entre le jour de réservation et le jour de départ

De plus, à partir de la variable date, plusieurs autres variables temporelles sont créées :
L'année de départ, le mois de départ, le jour de départ le jour de la semaine du départ (entre 1 = lundi et 7 = dimanche), la semaine de l'année du départ (entre 1 et 52) et le nombre de jour depuis le premier février 1970

Une base de données externe est aussi disponible contenant des données météorologiques. Dans un premier temps nous analyserons les résultats trouvés sur la base initiale, puis nous ajouterons la variable de température à celle-ci. Ensuite nous testerons quelques modèles sur l'ensemble des données disponibles incluant l'ensemble de la base externe. Par ailleurs, nous tenterons d'ajouter des variables qui semblent pouvoir apporter de l'information à notre modèle. Pour terminer, nous testerons de nouveaux algorithmes de régression avec la variable de température.

II. Analyse des modèles existants

Notre but principal est de créer un modèle qui minimise la RMSE moyenne, représentant la racine de la somme des erreurs au carré, ainsi que son écart type. Il existe déjà quatre modèles donnés en exemple : deux forêts aléatoires (avec et sans la variable de température aussi) et une régression linéaire (sans la variable de température).
Ces méthodes utilisées sont très différentes, étant donné que la Random Forest est non linéaire. Il peut donc être très intéressant de les comparer.
L'ajout de la variable supplémentaire de température ne semble pas permettre de diminuer la RMSE, mais seulement de réduire légèrement l'écart type, ce qui est plutôt pas mal.
Avec uniquement les variables de base :

- Arbre de régression : RMSE : 0.6277 +/- 0.0186
- Régression linéaire : RMSE : 0.6117 +/- 0.0149

Avec la variable de température en plus :

- Arbre de régression : RMSE : 0.6330 +/- 0.0108

Il faut noter que pour les deux forêts aléatoires, les hyper paramètres considérés sont les suivants : n_estimators = 10, max_depth = 10, max_features = 10.

A travers ces premier résultats, la régression linéaire semble faire moins d'erreurs lors de la prédiction du nombre de passagers comparé à la Random Forest regressor (nous avons aussi testé ce modèle avec la base contenant la variable de température). C'est donc le modèle le plus précis jusque-là, avec une RMSE de 0.61. Il faudra donc trouver l'algorithme qui fait mieux, ou bien les variables qui permettent de meilleures prédictions.

III. Ajout de variables supplémentaires

Dans un premier temps, nous avons cherché des variables disponibles de façon hebdomadaire et qui pourraient augmenter la précision du modèle. Etant donné que la régression linéaire est un modèle difficile à paramétrer davantage (bien que nous l'ayons tout de même testé), nous allons donc voir s'il est possible de mieux faire en paramétrant des arbres de régression.

3.1. Rendement du S&P 500

Tout d'abord, nous avons fait le choix d'ajouter une variable financière dans notre modèle, pour voir si celle-ci avait un impact sur le nombre de réservations. Nous utilisons le célèbre indice boursier américain qui est le S&P 500. Notre hypothèse est que celui-ci est un indicateur de la bonne santé financière du pays (et du monde même). Ainsi, lorsque le pays va bien financièrement, les gens partent plus en vacances, ce qui se traduirait par plus de passagers dans les avions. Nous calculons, via Excel, son rendement journalier et nous faisons une jointure sur nos données par la date de départ. Notre variable contient des valeurs manquantes pour les jours de départ où la bourse est fermée. Nous remplaçons ces valeurs manquantes par la valeur du rendement de la veille qui serait la meilleure estimation possible.
Sur cette base, et avec l'estimateur de Random Forest initialement construit (n_estimators = 10, max_depth = 10, max_features = 10), nous obtenons une **RMSE de 1.0101 et un écart-type de 0.0297.** En cherchant à optimiser ces résultats avec la GridSearchCV, la **RMSE obtenue est de 0.9943 avec un écart-type de 0.0302.**

Nous remarquons donc que l'ajout de cette variable empire toujours et considérablement nos modèles, et ce, malgré le tunning des hyper paramètres. Notre hypothèse semble être fausse, ou alors les américains ne partent pas en vacances en Amérique. Nous décidons de ne plus utiliser cette variable dans nos essais de modèles.

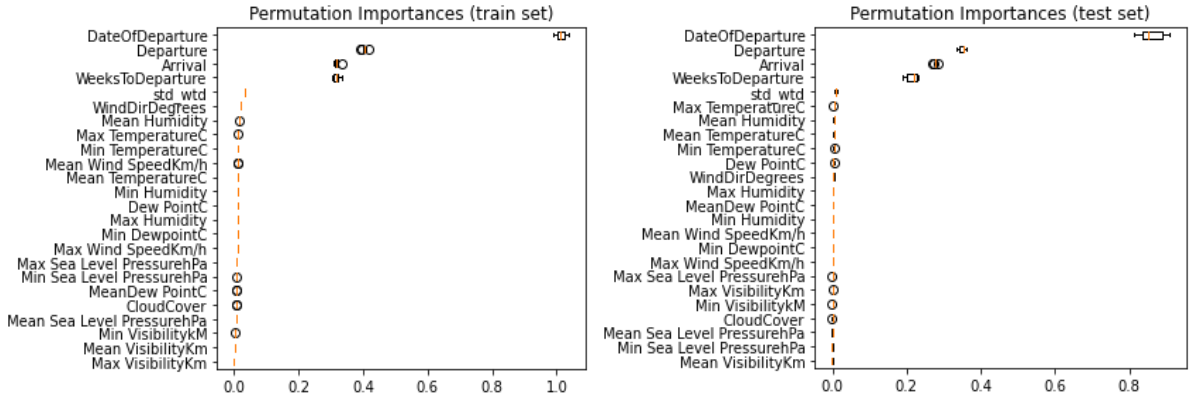
3.2. Cours du pétrole

Nous avons donc tenté d'introduire une autre variable financière : le cours du baril de pétrole. Notre hypothèse est la suivante : plus le pétrole est cher, plus le carburant des avions est cher et plus le billet coûte cher. Plus le billet coûte cher, et moins les personnes en achètent, soit car elles décident de prendre d'autres moyens de transport moins onéreux, ou alors parce qu'elles annulent leurs vacances. En incluant cette variables, les résultats n'ont pas été améliorés, mais plutôt détériorés tout comme avec le S&P 500. En effet, nous obtenons une **RMSE de 0.7946 et un écart type de 0.0194** avec la "meilleure" Random Forest sélectionnée.

Les variables ajoutées n'ont pas apporté plus de précision à nos prédictions. C'est la raison pour laquelle nous prenons la décision de travailler sur les données déjà disponibles.

IV. Amélioration des modèles déjà présents

Nous avons tout d'abord tenté d'améliorer les modèles déjà utilisés, simplement en réglant les hyperparamètres (pour RandomForest Regressor) à l'aide de la GridSearchCV() sur la base intégrant la variable de **température seulement**. Après plusieurs essais, le meilleur modèle donne une **RMSE de 0.5129 et un écart-type de +/- 0.0163**. Ces résultats sont les meilleurs trouvés en termes de RMSE jusque-là.
Les hyperparamètre utilisés sont les suivants :
bootstrap : True, max_depth : 50, max_features : 'auto', min_samples_leaf : 2, min_samples_split : 4, n_estimators : 200.



Nous avons aussi intégré plus de variables de météorologie à notre modèle, pour utiliser sur ces dernières des régressions linéaire, Ridge, Lasso et ElasticNet (il y a beaucoup plus de variables donc nous nous attendons au fait qu'une sélection de variables donne de meilleurs résultats qu'une régression linéaire seule) qui n'ont pas donnés de résultats intéressants. Nous revenons donc vers la Forêt Aléatoire établit sur ces données, et nous obtenons une **RMSE de 0.5440 et un écart-type de +/- 0.0127**. Les hyper paramètres correspondant à ces résultats sont les suivants : bootstrap : True, max_depth : 50, n_estimators : 200.

Que ce soit sur le train ou sur le test, l'importance des variables de météo reste très faible par rapport aux variables des modèles de base.

Tous nos résultats précédents ont démontré que seule la variable de température semble apporter une information pertinente à nos données de base. Cependant, il nous a semblé possible de pouvoir améliorer la RMSE à travers peut-être d'autres algorithmes.

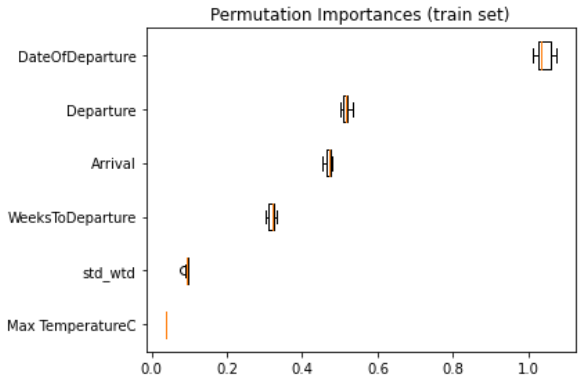
V. Introduction de nouveaux algorithmes

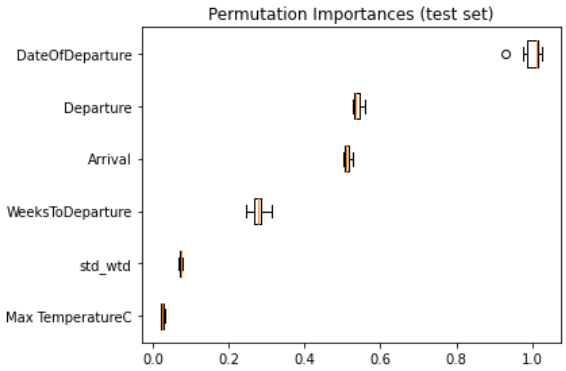
Nous avons donc essayé les KNN Regressor (les K plus proches voisins), qui n'ont pas donné de résultats convaincants. Pour ce faire, nous avons inclus l'ensemble des variables des données externes et naturellement supprimé les variables catégorielles. Cette méthode donne pour meilleur résultat une **RMSE de 0.9062 ainsi qu'un écart type de +/- 0.0238** avec les 20 plus proches voisins.

Pour terminer, comme jusqu'ici la Random Forest a été la plus efficace, nous avons décidé de mettre en place le Gradient Boosting Regressor. Celui-ci constitue les mêmes principes globaux que la Random Forest, à l'exception qu'il considère les mauvaises prédictions précédentes pour les améliorer ensuite.
Dans une première étape, nous avons lancé ce modèle sur la base incluant toutes les variables externes avec une Grid Search CV. Celle-ci donne de très bons résultats : une **RMSE de 0.4591 et un écart type de +/- 0.0197.**
Pour confirmer que la variable de prix du baril de pétrole n'est vraiment pas efficace, nous l'ajoutons aux variables externes et avons une forte dégradation des résultats comme attendu : **une RMSE de 1.0514 et une écart type de 0.025.**

Enfin, nous nous basons sur notre base initiale à laquelle est ajoutée seulement la variable de température. La Grid Search CV donne une **RMSE de 0.4010 et un écart type de +/- 0.0210.** Les

paramètres sélectionnés sont les suivants : loss = 'ls', learning_rate = 0.055, n_estimators = 2500, max_depth = 6.
A travers cet algorithme nous réussissons à obtenir un très bon **score de 0.851**. **Celui-ci sera donc notre modèle final correspondant au Test4 sur Ramp** (pseudo : Feriel) **où la RMSE est de 0.317**.
Ci-dessous l'importance des variables dans notre modèle :





Les mêmes variables arrivent au même niveau d'importance pour le train et le test ici. La date de départ semble influencer le nombre de passagers. Il semble logique qu'à certaines périodes de l'année il y ait plus de personnes qui prennent l'avion (les saisons estivales). En seconde position arrive l'aéroport de départ. En effet, il est très probable que ce soit les aéroports les plus grands qui font décoller les plus gros avions et donc le plus de voyageurs. L'aéroport d'arrivée serait aussi parmi les variables les plus importantes pour les mêmes raisons citées juste avant. Enfin, le nombre de semaines où la réservation est faite avant la date de départ impacte aussi le nombre de personnes voyageant.

Enfin, nous avons choisi de tester en plus, une régression Light GBM sur les variables de base et la variable de température. Dans un premier temps, nous lançons la régression sans aucun hyperparamètre et nous obtenons une **RMSE de 0.4308 et un écart type de 0.0177** !
Après optimisation à l'aide de GridSearchCV nous arrivons à descendre à une **RMSE de 0.3922 et un écart type de 0.0200**. Voici les hyperparamètres utilisés :
learning_rate : 0.1, n_estimators : 450, num_leaves : 40.
Le **score** reste aussi très bon avec une valeur de **0.8477**. Ce modèle est donc en concurrence avec le Gradient Boosting trouvé précédemment (avec une RMSE de 0.320 pour la submission sur Ramp).
Ci-dessous nous remarquons que l'importance des variables est la même que pour le Gradient Boosting précédent.