Projet de Deep Learning : Estimation du nombre de passagers aérien

Yanis Ahdjoudj, Lucas Diaz

February 10, 2021

1 Exposition

La pandémie du Covid-19 a fait chuter de 60 % le nombre de passagers des compagnies aériennes dans le monde en 2020, selon l'Organisation de l'aviation civile internationale (OACI) des Nations-unies. Le nombre de passager par vol est très important car, le connaissant, il permet de veiller à la bonne répartition des avions selon les itiniéraires plus ou moins chargés. Il est donc important et intéressant de connaître ce nombre de passagers selon la date ainsi que la destination.

Nous allons prédire un champ appelé logPax lié au nombre de passagers par vol à l'aide de données provenant de réservations de billets d'avion provenant d'une société anonyme. Ces données regroupent le cadre spatio temporel du vol ainsi que des variables sur le moment de réservation. On a donc une base de données dans laquelle on peut exploiter les aéroports et les dates des vols. Cela concerne des vols intra muraux aux Etats Unis, sur une période allant de 2011 à 2013. Nous aggrémenterons cette base à l'aide de données extérieures. Nous mesureront la qualité de notre prédiction à l'aide du score de la RMSE.

2 External Data

Pour mieux estimer ce nombre de passagers, on recherche de nouvelles données qui auraient une corrélation avec notre cible pour les intégrer dans notre pipeline.Les données récupérées vont se faire selon deux clés de jointure, les aéroports et les dates.

Concernant les données temporelles, on commence par intégrer les données externes fournies par la plateforme Ramp. On récupère les données météorologiques sur la température et les intempéries.

Pour estimer le marché du transport aérien, on récupère les cotations boursières de American Airlines sur la période (Yahoo Finance).

Pour inclure une variable pouvant témoigner du contexte économique, on intègrera dans notre modèle le prix du fuel récupéré sur le site de l'U.S. Energy Information Administration.

On intègre également les informations sur les jours fériés et périodes de vacances.

Concernant les aéroports, on va se référer aux villes et aux comtés qui englobent ces aéroports. A l'échelle des villes, on récupère les populations. Puis à une granularité plus importante, on récupère les salaires médians par comté (The Department of Numbers). Finalement notre liste de variables utilisées est la suivante :

- # Variables de base

- **Departure**: Aéroport de départ
- **Arrival** : Aéroport d'arrivée
- **WeeksToDeparture** : Temps entre la réservation et le départ (en semaine)
- **std wtd** : Ecart type de WeeksToDeparture
- **temperature*** : Température locale

- # Variables ajoutées

- **precipitation** * : Présence de brouillard, pluie, neige orage...
- **wage median*** : Salaire médian pour le comté où se trouve l'aéroport
- **beach*** : Présence d'une plage
- **passenger per year*** : Nombre de voyageurs moyen par an par aéroport
- **population*** : Population de la ville où se trouve l'aéroport
- **distance** : Distance entre les deux aéroports
- **diff temp** : Différence de température entre l'arrivée et le départ
- **score*** :
- **Open** : Indice boursier d'American Airlines
- **prix** : Prix du fuel au Etats Unis
- **is holiday** : Période de vacances
- **year** : Année du départ de vol
- **month** : Mois du départ de vol
- **day** : Jour du départ de vol
- **weekday** : Jour de la semaine du départ de vol (Lundi:0, Mardi:1...Dimanche : 6)
- **weekend** : Variable catégoriel, Jour de weekend

(* Variable présente à la fois pour l'aéroport de départ et l'aéroport d'arrivée)

3 Modélisation

Pour la modélisation nous utilisons pour terminer notre pipeline un modèle de gradient boosting fourni par la librairie scikit-learn. Le choix des hypermaramètres se base sur les multiples tentatives que nous avons effectuées, il en résulte le modèle suivant.

```
# Regressor to do the prediction
regressor = GradientBoostingRegressor(loss='ls', learning_rate=0.01,
            n_estimators=2000, subsample=1.0, criterion='friedman_mse',
            min_samples_split=10, min_samples_leaf=5,
            min_weight_fraction_leaf=0.0,max_depth=4,
            min_impurity_decrease=0.0, min_impurity_split=None)
```

Figure 1: Modèle utilisé dans la pipeline

Son exécution plus en détail et le fichier estimator utilisé pour Ramp peuvent être trouvés sur le repo GitHub de notre projet **ici**

4 Résultats

Nous obtenons au final les résultats suivants à partir de la soumission sur Ramp :

Avec une RMSE de 0.301 nous sommes satisfaits des résultats

team	submission	rmse	train time [s]	validation time [s]	max RAM [MB]	submitted at (UTC)
Yanis_Ahdjoudj	Rikudo	0.301	1609.372521	379.986788	0.0	2021-02-10 22:36:42

Figure 2: Résultats du modèle estimé via la platforme RAMP

5 Interprétation

Dans cette section nous interpréterons les résultats par la méthode des valeurs de Shapley.

Cette méthode a été développée en théorie de jeux par en Lundberg et Lee en 2016. Elle va nous permettre d'étudier les variables les plus importantes dans le modèle mais également leur sens sur la variable cible (positif ou négatif).

Il consiste à moyenner l'impact d'une variable a sur notre cible pour toutes les combinaisons de variables possibles.

Le graphique ci-après nous renseigne, en plus de l'ampleur, sur le sens d'agissement sur le nombre de passagers.

Chaque point indique une valeur SHAP et mesure l'impact sur la valeur prédite (positif ou négatif). La couleur de ce point indique si la variable prend une valeur faible ou élevée (par rapport à elle-même).

On peut voir que plus on se rapproche du weekend, plus le nombre de passagers sera élevé en moyenne. Les vols les plus longs et à plus grande distance font baisser le nombre de passagers. Que ce soit pour le départ ou la destination, une importante population de la ville induit en moyenne une hausse du nombre de voyageurs. La lecture du graphique sur la variable 'WeektoDeparture' nous indique que plus on reserve notre vol en avance , et plus il y aura de passagers au sein de ce vol. L'interprétation de 'std wtd' nous indique que plus il y a des différences dans les temps de reservations, moins le nombre de passagers sera considérable. D'après la lecture de 'passenger per year arrival', on déduit logiquement

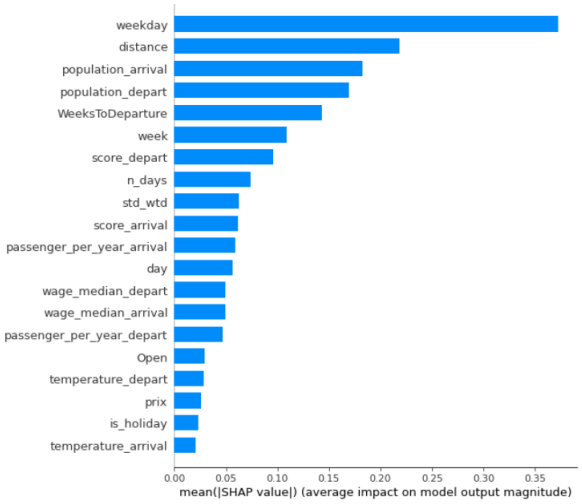

Figure 3: Importance des variables par l'interprétation des valeurs de Shapley

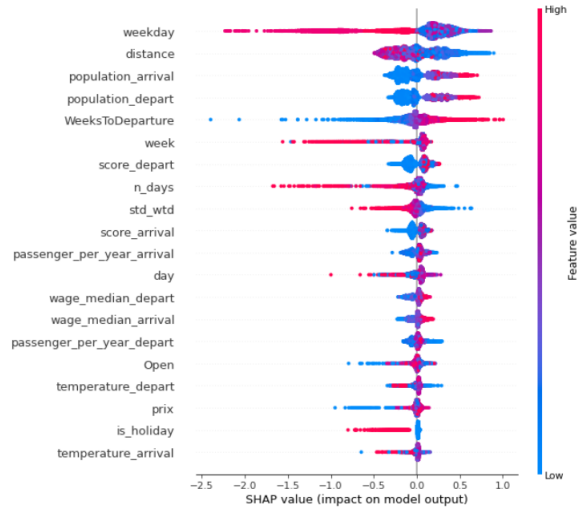

Figure 4: Interprétation du sens des variables par les valeurs de Shapley

que plus l'aéroport est fréquenté, plus le nombre de voyageurs est important. On peut voir aussi que plus les salaires sont élevés dans le comté et plus il y a de l'affluence sur les vols, cela est aussi bien vrai pour le départ que l'arrivée. On voit aussi avec 'Open' que plus la côte des action de American Airlines est basse et moins les passagers sont nombreux ; d'où l'importance de prendre en considération le contexte économique du secteur. De même plus le prix du fuel est bas et moins les voyageurs seront nombreux. Enfin les périodes de vacances sont synonymes en moyenne de peu de passagers par vol.