Hotel booking demand

1 - Introduction

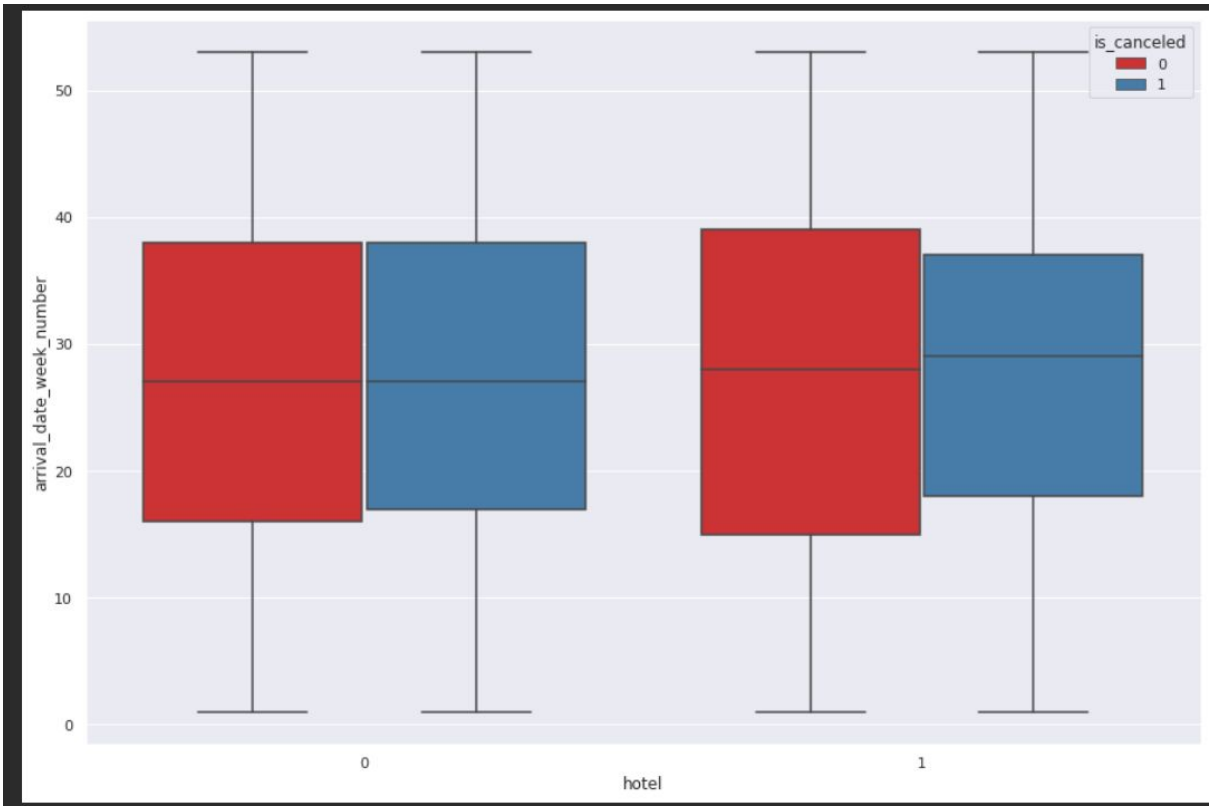
Le jeu de données à étudier représente des réservations d'hôtel.
Les hôtels concernés sont 2 hôtels situés au Portugal, un hôtel de type urbain et un hôtel de type divertissement.
Les données présentent s'étalent du 1er Juillet 2015 au 31 Aôut 2017.
31 données sont représentées, telles que les dates d'arrivée et de départ, le nombre de nuits passées, le nombre de clients, les annulations etc...
L'objectif ici est de construire un modèle d'apprentissage machine capable de prédire les probabilités d'une annulation
pour une commande donnée.

2 - Analyse des données

Observons le taux d'annulations en fonction du type d'hôtel concerné :

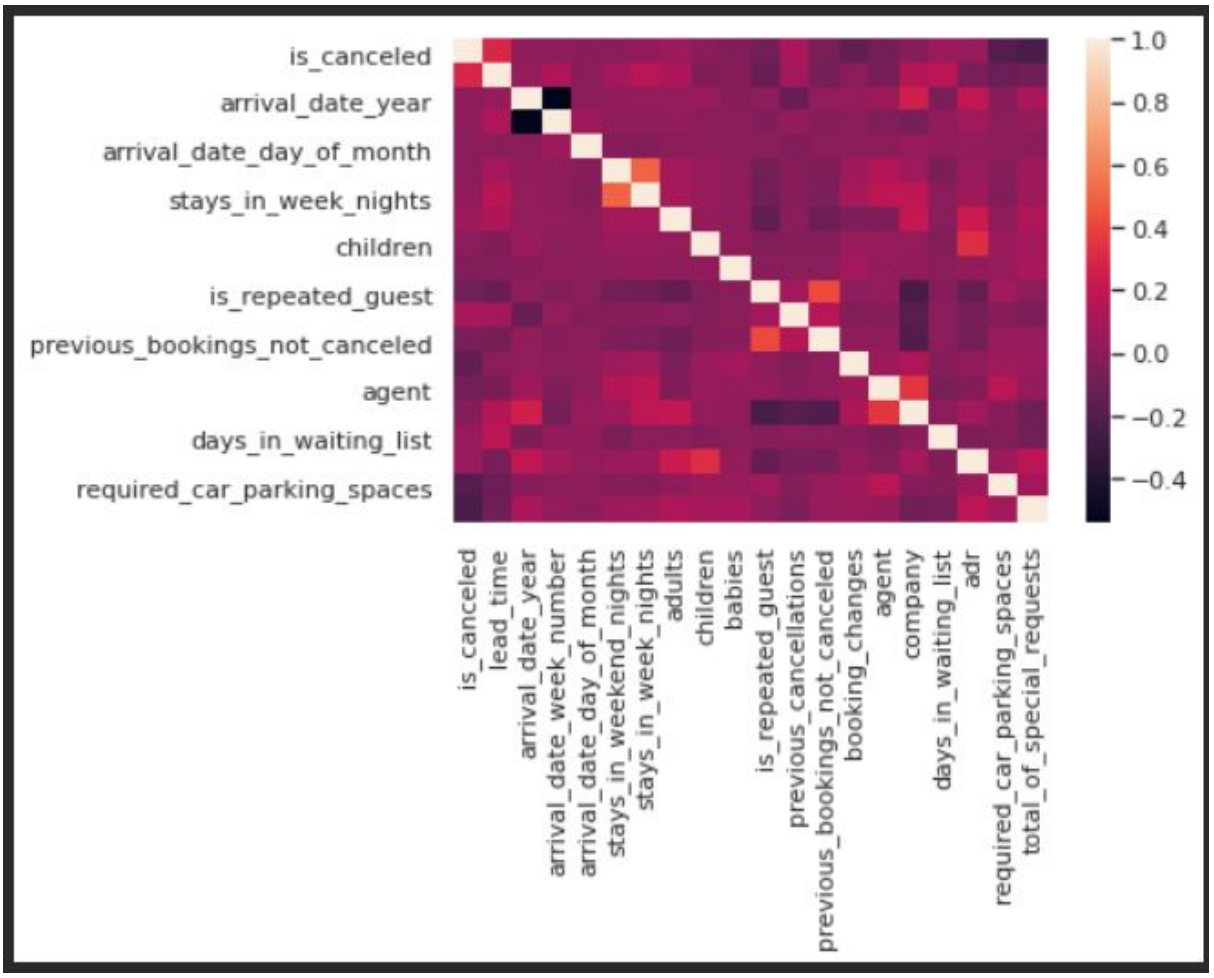
On peut voir sur ce diagramme que les hôtels urbains sont beaucoup plus sujets aux annulations que les hôtels de divertissements. Ce sont aussi ces hôtels qui cumulent le plus de réservations totales.

Intéressons-nous maintenant à la répartition du type d'hôtel en fonction de la semaine d'arrivée en observant l'annulation.



Les réservation annulées pour les hôtels de divertissement (rouge) sont plus importantes et réparties sur une plus longue période de l'année que pour les hôtels urbains qui ont un taux d'annulation réparti sur une période plus courte. On n'observe ici aucune valeur aberrante particulière.

Analysons maintenant les corrélation entre les variables au travers d'une matrice de corrélation :



Cette matrice ne montre ici aucune corrélation particulière entre nos données.

2 - Construction du modèle

Afin de trouver le modèle d'apprentissage le plus optimisé pour notre prédiction, une pipeline a été utilisée afin de tester différents jeux de paramètres pour les algorithmes.
Le résultat de cette opération nous montre que l'algorithme idéal est l'algorithme des plus proches voisins avec l'hyperparamètre "k_neighbors" à 1.
Le score d'accuracy est de **90%**.

Nous avons essayé d'améliorer ce score en retravaillant nos données.
Pour cela certaines colonnes du jeu de données ont été regroupées.
Les colonnes

- `stays_in_weekend_nights`
- `stays_in_week_nights`

ont été regroupées en une même colonne "`stays_in_week_nights`".

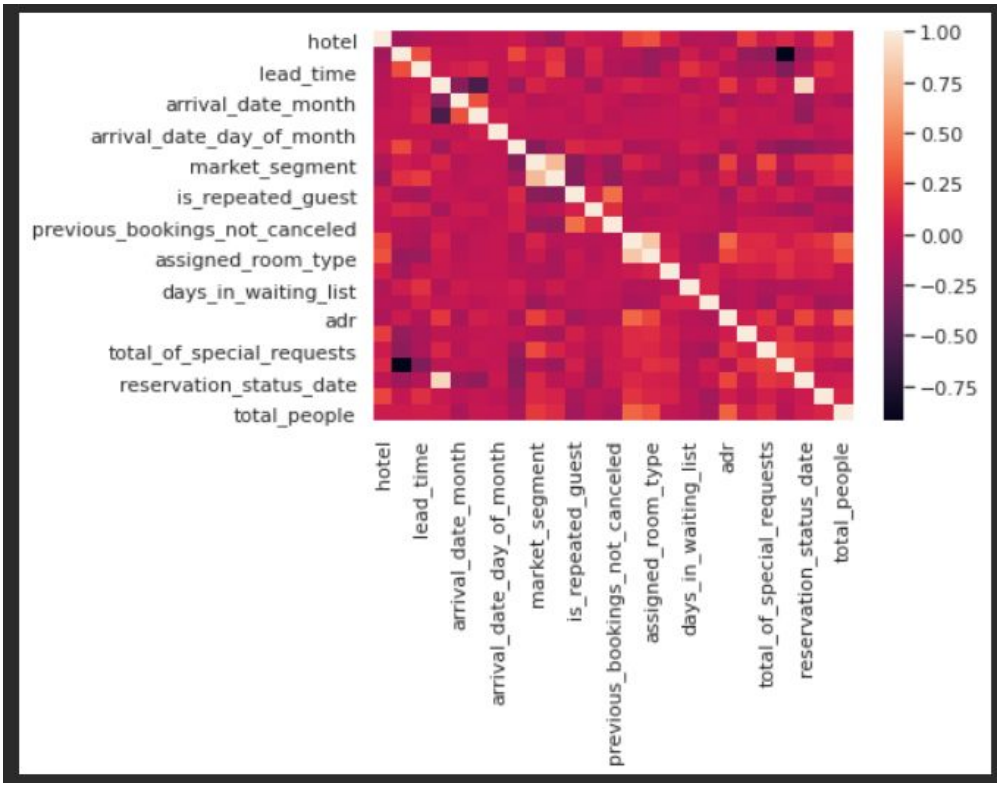
Les colonnes

- `adults`
- `children`
- `babies`

ont été regroupées en une même colonne "`total_people`".

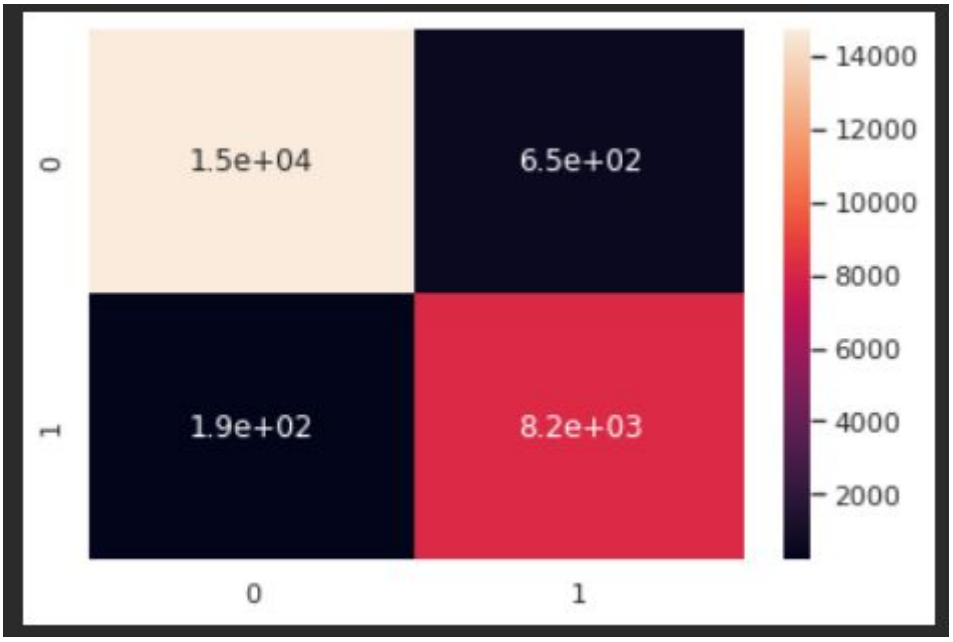
Nous avons ensuite supprimé du jeu de données les colonnes ainsi regroupées.

En réanalysant les corrélations à l'aide d'une matrice de corrélation on observe qu'aucune corrélation significative n'est apparue :



En revanche, après calcul du score d'accuracy avec les nouvelles données, l'accuracy est passée de 90% à 96%.

Dans la matrice de confusion suivante on obtient plus de précision sur les erreurs commises par le modèle.
On observe ici un taux de faux positifs et de faux négatifs égal et très bas, ce qui confirme la performance du modèle.



Conclusion

On est donc parvenu à obtenir un modèle de prédiction performant pour notre cible. La prédiction des annulations de réservation peut permettre au métier de prévoir avec plus de précision les ressources humaines et matérielles dont ils auront besoin sans avoir de surplus que pourrait engendrer un trop grand nombre d'annulation.
De plus, les hôtels pourraient envisager de louer un surplus de chambres en fonction des annulations prédites et ainsi optimiser leur rendement.