Prix des locations à Rio

Rapport projet python

Daucourt Alex

Harle Remy

Maron Elisa

29 janvier 2023

Introduction 3

Pré-traitement des données 3

I. Gamme de prix 3

II. Caractéristiques des logements 4

III.Localisation des logements 4

Modèle de prédiction 6

I. Avec le nom de l'annonce 6

II. Comparaison des modèles 7

III. Interprétabilité du meilleur modèle 7

Conclusion 14

Introduction

Airbnb est une plateforme qui permet aux gens de mettre en location, de trouver et de louer des maisons de vacances, des appartements et d'autres hébergements. C'est une façon pour les propriétaires de gagner de l'argent en louant leurs propriétés aux voyageurs, et c'est une façon pour les voyageurs de trouver des endroits uniques et abordables où séjourner pendant leurs vacances. Airbnb a été fondé en 2008 et est depuis devenue l'une des plateformes de location de vacances les plus populaires et les plus réussies au monde.

Lorsqu'un utilisateur décide de mettre en location un de ses biens sur la plateforme, Airbnb souhaite proposer au propriétaire un outil de calcul pour prédire le prix potentiel du bien proposé. L'utilisateur disposera de la classe de prix dans laquelle se situe son bien. Ces classes seront définies en fonction du marché immobilier de la région. Ce projet est réalisé sur la base de 35 602 logements à Rio au Brésil.

Pré-traitement des données

Nous disposons de plusieurs informations sur les logements en location. Tout d'abord nous avons accès au type de bien, ainsi qu'à la capacité d'accueil maximale du logement et au nombre de chambre, de salle de bain et de lit à disposition. Ensuite nous avons le nom du quartier dans lequel se trouve le bien mais aussi sa position exacte avec sa longitude et sa latitude. De plus, nous connaissons le titre de l'annonce du logement en location choisi par le propriétaire. Enfin nous disposons du prix des appartements déjà en location afin d'entrainer nos modèles.

I. Gamme de prix

Nous avons accès aux prix de chaque logement par nuit (en $). Afin de créer nos classes, nous calculons les différents quantiles des variables et les prendrons comme valeur seuil.

Quantille	Prix
20 %	129 $
50 %	300 $
80 %	790 $

Ainsi la classe 0 regroupera les logements à un prix inférieur à 129$, la classe 1 les logements entre 129 $ et 300 $, la classe 2 entre 300 $ et 790 $ et enfin la classe 3 les prix supérieurs à 790 $.

II. Caractéristiques des logements

Cette première partie sera consacrée aux types du bien proposé ainsi qu'a ces différentes pièces et sa capacité d'accueil. Nous commençons par traiter les valeurs manquantes.

Variable	Nombre de valeur manquante
Bedrooms	24
Bathrooms	69
Beds	49

Comme nous disposons de beaucoup de données, nous avons choisi de remplacer les valeurs manquantes par la médiane respective de chaque variable.
Nous nous intéressons ensuite au logement type que l'on trouve à Rio. En moyenne les logements disposent de 2 chambres, 3 lits et 2 salles de bain pour une capacité maximal de 4 personnes.
Nous disposons également de différents types de logement :

Type de logement	Nombre de logement
Logement entier	24929
Chambre privée	9819
Chambre partagée	854

III. Localisation des logements

Cette partie concerne la situation géographique des différents logements. Pour 3030 logements nous n'avons pas le nom du quartier, on le remplace par une chaine de caractère vide. En revanche nous disposons de la longitude et de la latitude qui nous permet d'accéder à la position exacte du bien. Avec une fonction il est possible de trouver le quartier du bien en fonction de ces données. Cependant cette fonction est très couteuse en temps et ne retourne parfois aucune valeur. De plus, certains quartiers ne correspondent pas à ceux dont nous disposons dans la base de données. Nous avons

donc décidé de ne pas remplacer ces valeurs et de les laisser de coté lors de la modélisation, car la base de données est suffisamment grande.
Cependant nous utilisons les coordonnées des différents logements pour déterminer la distance avec certains lieux marquants, comme le stade du Maracana et la statue du Christ rédempteur. Nous avons créé deux nouvelles variables avec la distance en kilomètre entre le bien et ces deux monuments. Grace à la position des logements nous avons également créé une carte nous permettant de voir la gamme de prix par logement.

Gamme de prix des logements en location à Rio

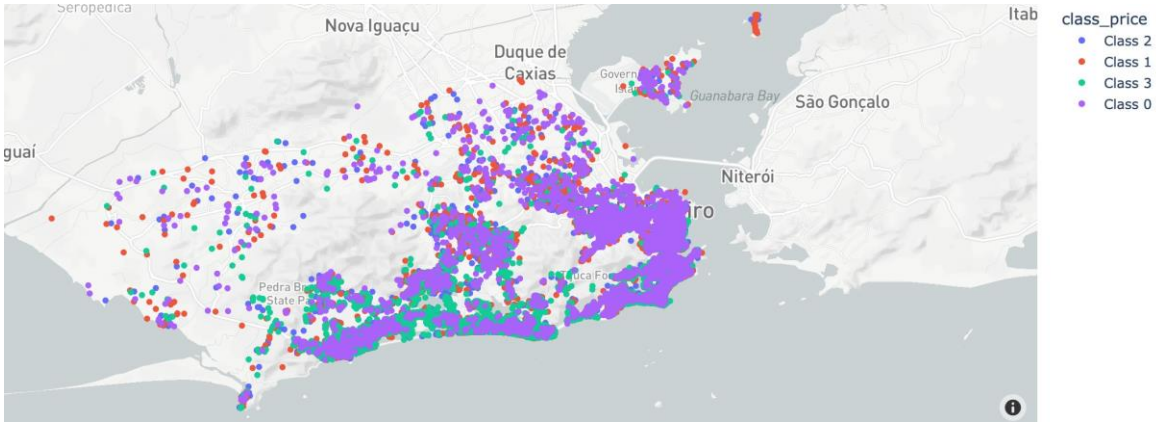


Figure 1.1 : Cartographie des logements par gamme de prix

Modèle de prédiction

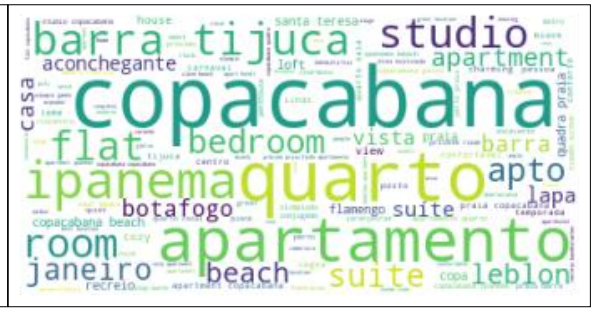
I. Avec le nom de l'annonce

Nous avons regardé si l'intituler de l'annonce avec des détails sur le logement pouvait être utilisée pour prédire la classe du prix de celui-ci.

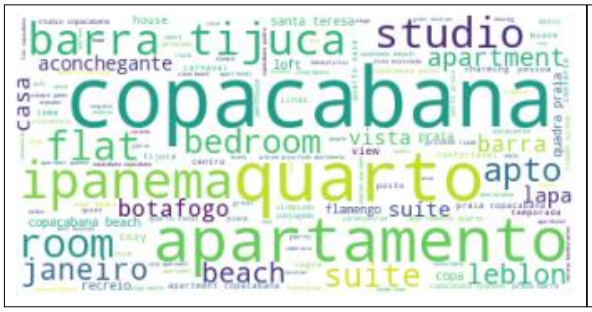
Pour chaque gamme de prix, nous regroupons les mots les plus utilisés pour la description.

Mots les plus utilisés pour décrire les biens de la classe 0

Mots les plus utilisés pour décrire les biens de la classe 1



Mots les plus utilisés pour décrire les biens de la classe 2



Mots les plus utilisés pour décrire les biens de la classe 3

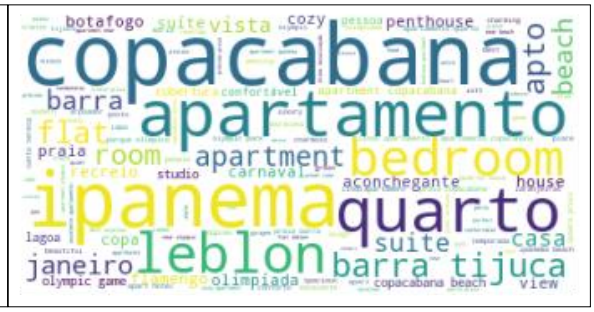


Figure 2.1 : Nuages de mots les plus répandus dans le titre des annonces

Les nuages de mots pour décrire les différentes classes sont assez semblables. Il n'est donc pas pertinent de créer un modèle de prédiction en utilisant ces mots car il ne donnerait pas de bons résultats. Ce modèle pourrait être amélioré en retirant les mots contenant les noms des quartiers et les mots du type « room », « casa » ou « studio » et tenir compte les autres données à disposition.

II. Comparaison des modèles

Nous allons implémenter différents modèles, chercher leurs paramètres optimaux puis les comparer afin de faire des prévisions avec le meilleur modèle obtenue. On remplace les variables qualitatives en quantitatives par indexation. On divise ensuite notre échantillon en deux, 70 % de notre échantillon sera pour entrainer le modèle, et les 30% restant pour la validation. On compare 100 modèles différents par validation croisée avec le score accuracy.

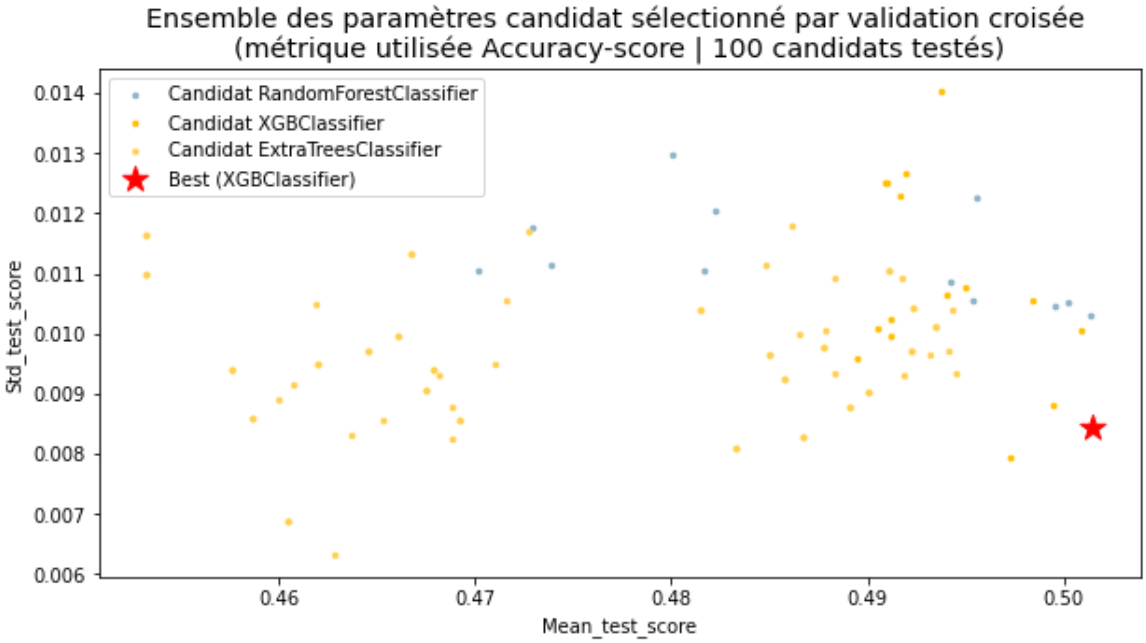


Figure 2.2 : Comparaison des modèles

Le meilleur modèle est obtenu avec la méthode de gradient boosting. On fera donc nos prédictions avec le modèle XGB avec les paramètres :

-learning_rate : 0.1

-max_depth : 7

-n_estimators : 50

On peut alors passer à la prévision.

III. Interprétabilité du meilleur modèle

Nous regardons maintenant quelles sont les variables qui ont le plus d'impact sur nos prévisions et de quelle façon elles permettent de prédire le prix d'un logement.

Importance des variables

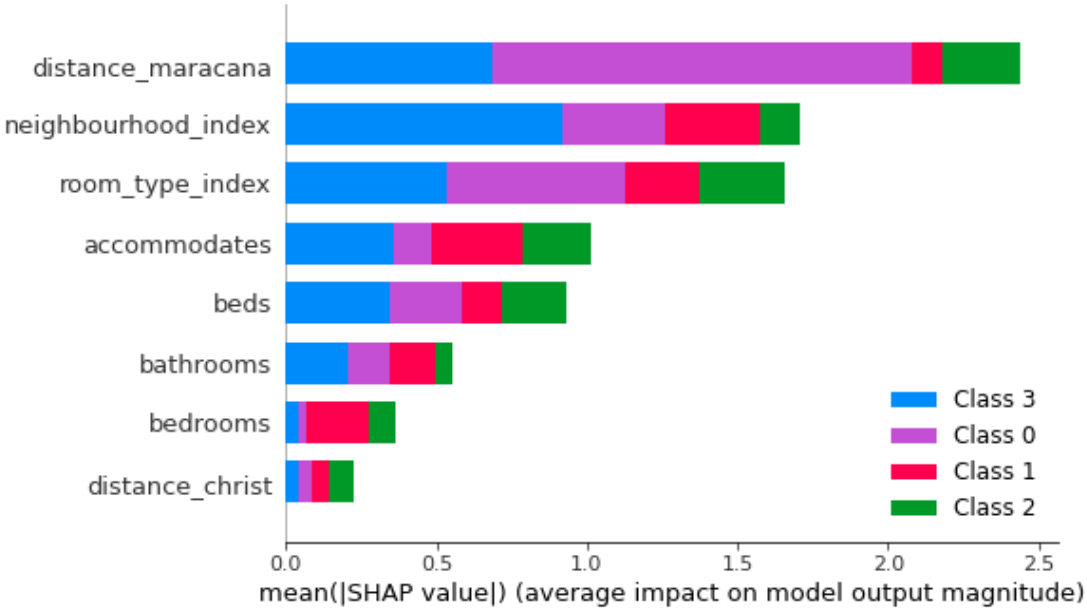
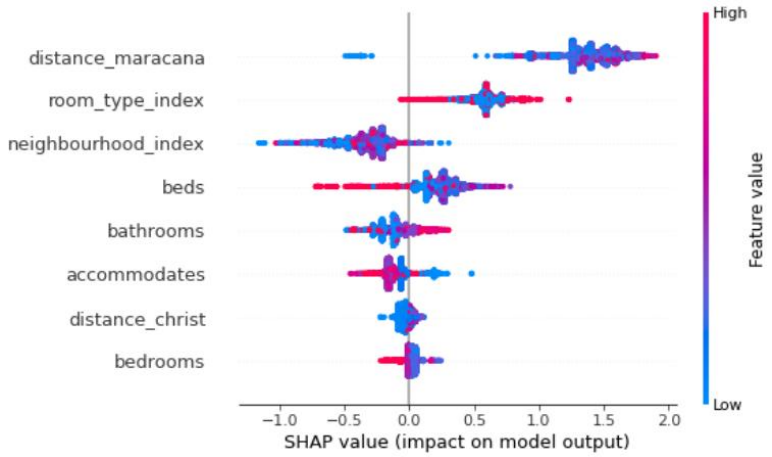


Figure 2.3 : Importance des variables

Nous voyons sur cette figure l'importance des variables pour chacune des classes. Pour ce modèle par exemple, la variable qui a le plus d'impact pour la classe 0 est la distance au Maracana. Pour la classe 3, c'est la variable "neighbourhood" qui correspond au quartier. Le type de logement et la capacité maximale d'accueil sont également des paramètres importants pour déterminer le prix d'un bien pour toutes les classes.

Si on veut plus de précisions concernant une classe en particulier, on peut l'afficher, comme sur ce graphique pour la classe 0 :



L'axe des ordonnées reprend les variables explicatives par ordre décroissant d'importance comme pour le précédent tracé.
L'axe des abscisses représente les valeurs SHAP. Une valeur négative indique une relation négative et une valeur positive indique une relation positive de la variable explicative (ici, l'appartenance à la classe 0). La barre de couleur sur la droite indique si la variable est élevée (en rouge) ou basse (en bleu). Par exemple, on voit que pour la

variable "beds" indiquant le nombre de lits, un nombre élevé de lit diminue la probabilité d'appartenir à la classe 0.

Intéressons-nous maintenant aux dependence plot, en particulier ceux de la dépendance de la variable distance Maracana pour les 4 différentes classes.

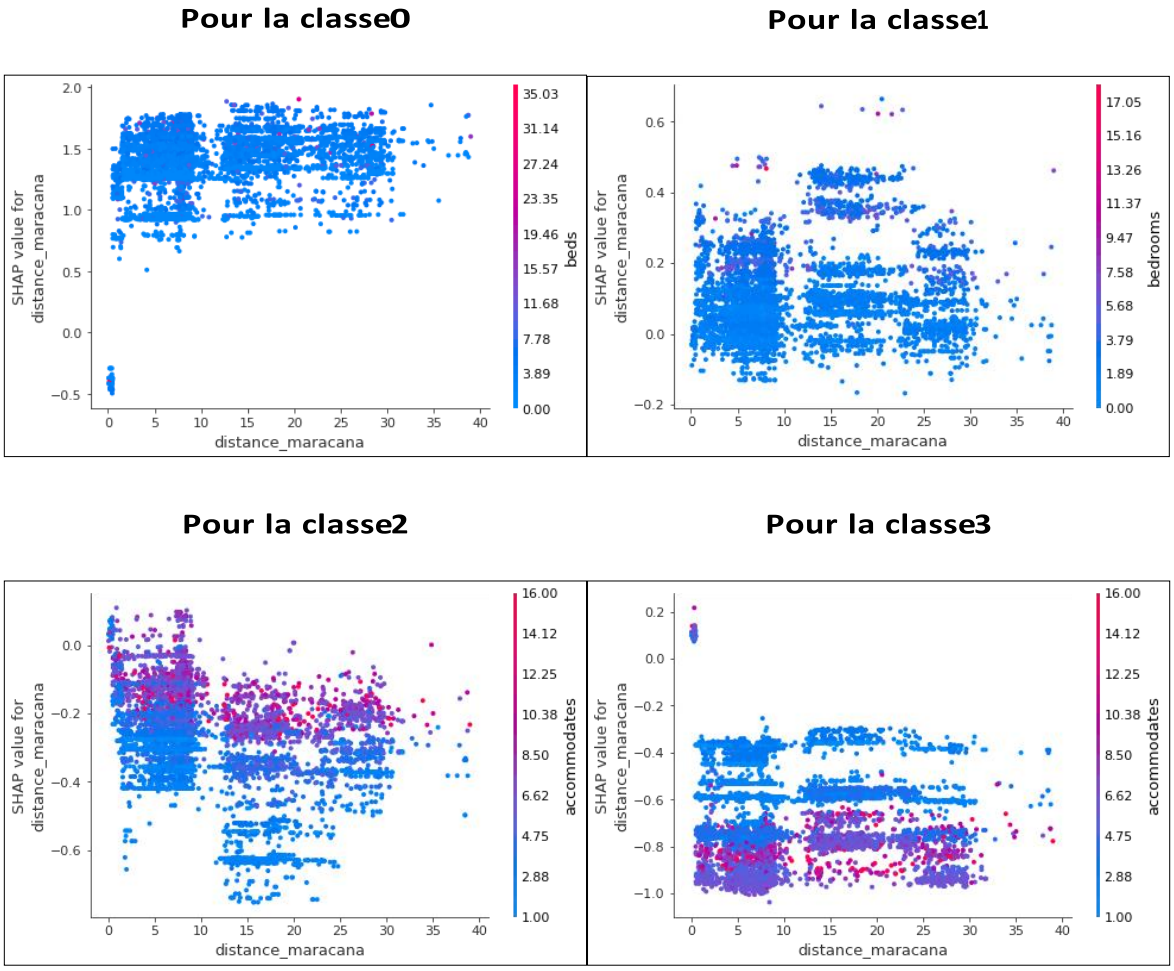


Figure 2.4 : Dépendance de la distance Maracana avec le logement

L'axe des ordonnées représente les valeurs SHAP de la variable choisie, quant à lui l'axe des abscisses représente les valeurs initiales de cette même variable.
La barre de couleurs sur la droite représente une échelle pour les valeurs de la variable qui serait le plus en interaction avec la variable 'distance Maracana'.
Les points à l'intérieur du tracé représentent l'impact global conjugué des deux variables dans la prédiction de l'appartenance à une classe.
Par exemple, pour la classe 0, on voit qu'une distance inférieure à ~2km du Maracana diminuerait la probabilité d'appartenir à la classe (car le SHAP est inférieur à 0). Au contrairement, à plus de 2km, la probabilité d'appartenir à cette classe augmente. De plus, la variable d'interaction choisie (variable 'beds') ne semble pas avoir une interaction forte qu'on peut interpréter, car la quasi-totalité des points sont bleus.

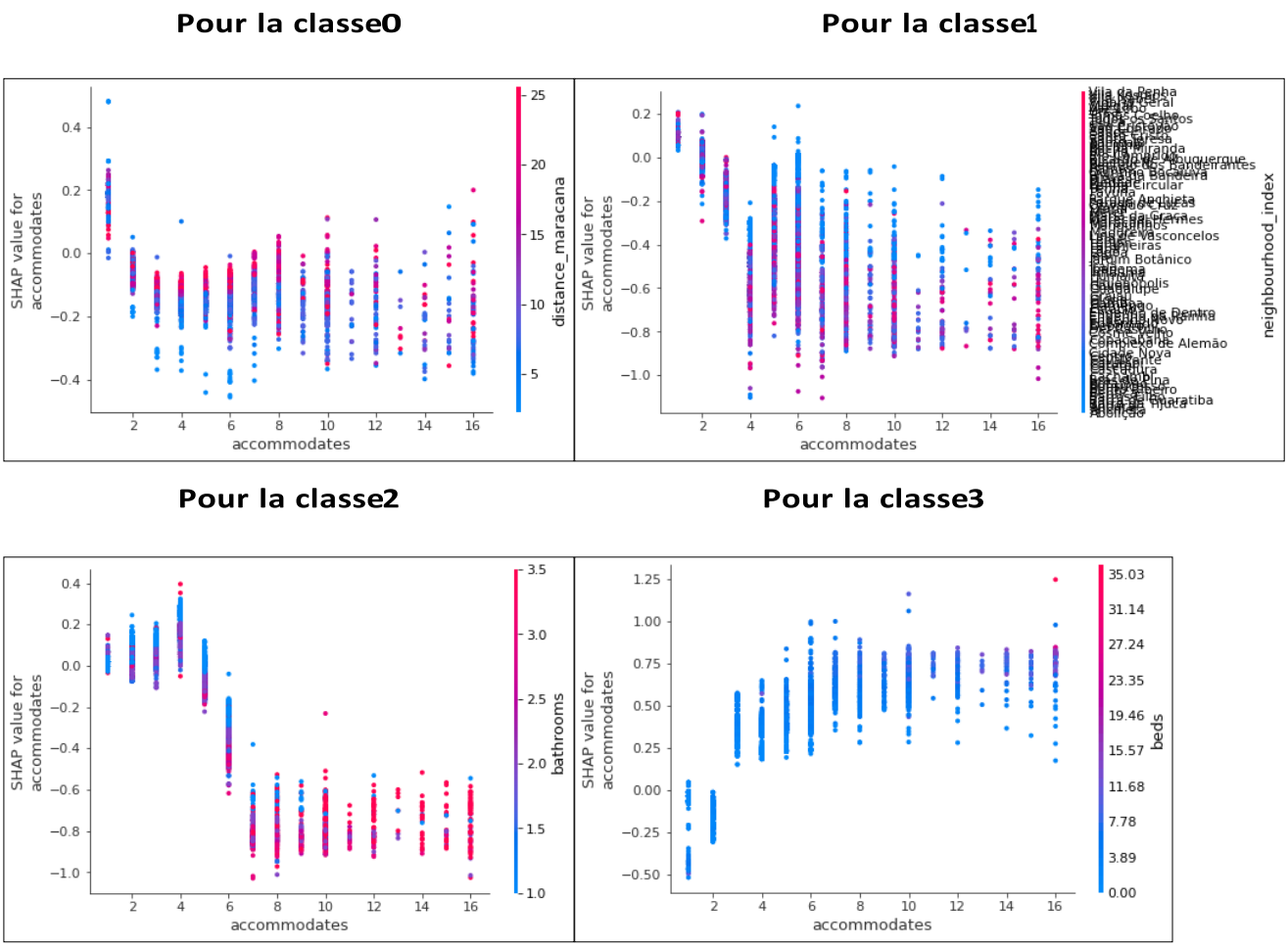


Figure 2.5 : Dépendance de la capacité maximale du logement

Passons maintenant aux dependence plot de la variable accommodates.
Pour la classe 0, on remarque que seulement une capacité d'accueil à 1 personne a une valeur SHAP positive. Cela signifie qu'un faible capacité d'accueil augmente les chances d'être dans la classe 0. Au contraire, à partir de 2 personnes, la probabilité d'être dans la classe 0 diminue.

Si on utilise les explainer.expected_value, on peut voir les base value pour chaque classe. On peut ensuite déterminer les probabilités en moyenne pour un appartement d'appartenir à chaque classe. En fait, on se rend compte que ces probabilités sont logiquement égales aux répartitions des classes dans notre échantillon total.

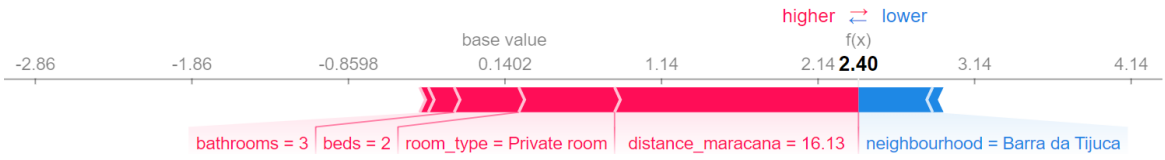
On peut calculer les probabilités de cette manière :

$$P(Classei) = \frac{\exp(expectedvalue_i)}{\sum expectedvalue}$$

	Explainer.explected.value	Softmax
Classe 0	0.14023232	0.17
Classe 1	0.7458794	0.31
Classe 2	0.73057294	0.31
Classe 3	0.2511787	0.19

Maintenant, nous allons analyser un force plot. Pour ce faire, nous allons sélectionner un Airbnb de façon aléatoire et observer les 4 force plot correspondant aux 4 classes.
Les valeurs sur l'axe sont les valeurs SHAP représentées sous forme de ln(odds).
La valeur base value est la prédiction moyenne dans notre jeu de données, celles qu'on vient de calculer. Analysons maintenant le premier force plot :

Pour la classe 0

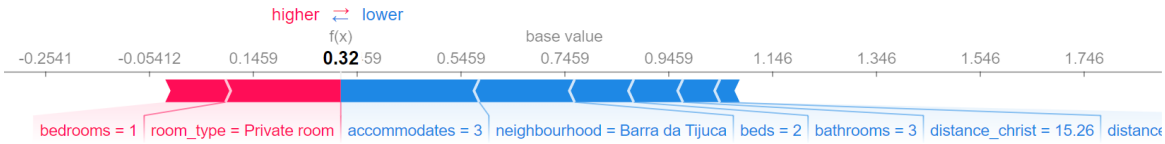


La valeur output value est de 2.40. C'est la prédiction d'être dans la classe 0 par notre modèle pour le Airbnb numéro 66 de notre échantillon test.
La couleur rouge représente les variables explicatives poussant vers le haut la prédiction : on voit que le nombre de salles de bain, de lit, le type du Airbnb et la distance du Maracana augmentent la probabilité d'être dans la classe 0. Au contraire, la couleur bleue représente les variables explicatives poussant vers le bas, le quartier et la capacité maximale.
On peut calculer la probabilité en utilisant cette formule :

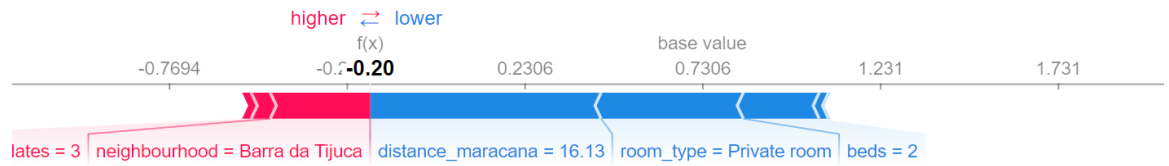
$$P(x) = \frac{e^{ln(odds)}}{\sum e^{ln(odds)}}$$

On trouve une probabilité de 77,9% que cet appartement appartienne à la classe 0. De manière analogue, on obtient une probabilité de 9,7% d'appartenir à la classe 1, de 5.8% d'appartenir à la classe 2 et finalement une probabilité de 6.5% d'appartenir à la classe 3.

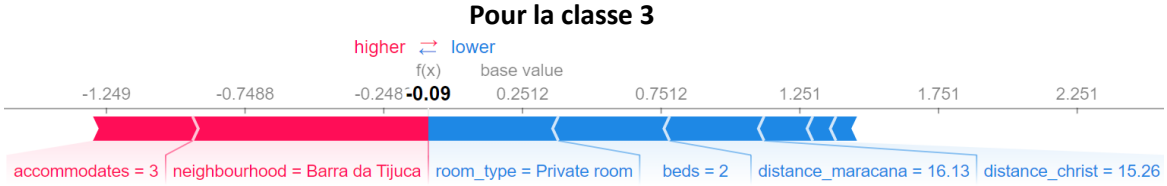
Pour la classe 1



Pour la classe 2



Pour la classe 3



On peut vérifier la classe réelle de notre appartement en allant le chercher dans notre échantillon de base, et en effet on constate qu'il fait partie de la classe 0.

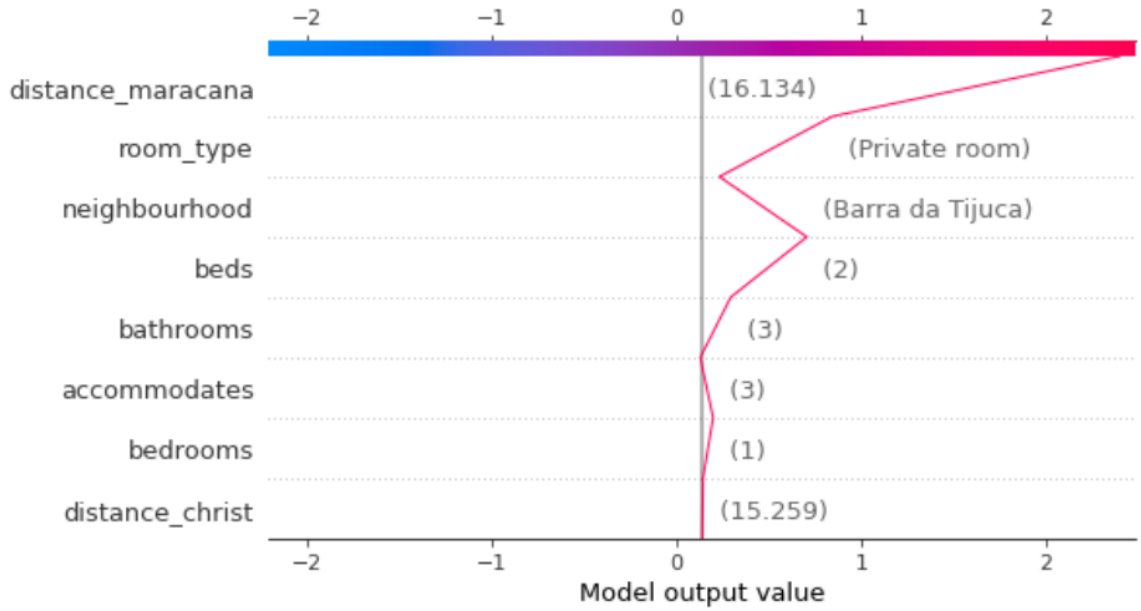
Caractéristique du Airbnb

id	2101665
host_id	7955669
room_type	Private room
neighbourhood	Barra da Tijuca
accommodates	3
bedrooms	1.0
bathrooms	3.0
beds	2.0
price	110.0
name	quarto confortável
latitude	barra da tijuca
longitude	-23.006641
class_price	-43.348852
distance_christ	Class 0
distance_maracana	15.258985
	16.134081

Pour aller plus loin, nous pouvons regarder les décisions plot. Les décision plot de SHAP montrent comment les modèles complexes arrivent à leurs prédictions (c'est-à-dire comment les modèles prennent des décisions).

Un décision plot peut révéler comment les prédictions changent au fur et à mesure de la prise en compte des variables explicatives.

Décision plot



Par exemple toujours avec le même Airbnb, on voit que l'algorithme par du bas du graphique avec la base value 0.2512 et va update ça shape value en fonction des caractéristiques du Airbnb jusqu'à arriver à la valeur 2.40 en haut du graphique que l'on a déjà vu précédemment dans le force-plot.

Conclusion

Pour améliorer la performance du modèle, on pourrait grouper certains quartiers avec des caractéristiques communes. Par exemple les quartiers le long de la plage, ceux avec un aspect historique ou bien avec des activités touristiques incontournables. On pourrait également prendre en compte la proximité avec l'aéroport et les autres moyens de transport. Il est également possible d'identifier les biens atypiques ou en faible quantité pour déterminer un niveau de rareté du logement. Il y a aussi d'autres pistes d'amélioration concernant les valeurs manquantes, l'analyse des titres (avec une meilleure liste de stopword) et de pousser plus loin l'interprétabilité du meilleur modèle.