Reporte Ciencia de los Datos

Chazelas Agathe
Conseil Eloïse

Sommaire

Introduction et objectif

La plateforme Airbnb aimerait prédire les prix des logements dans certaines villes des Etats-Unis. Pour ce faire, nous avons à notre disposition un échantillon de données. Pour une entreprise comme Airbnb, il peut être pratique de pouvoir prédire les prix des logements pour que les hôtes soient compétitifs entre eux et pour que les voyageurs aient une idée du prix immédiatement.

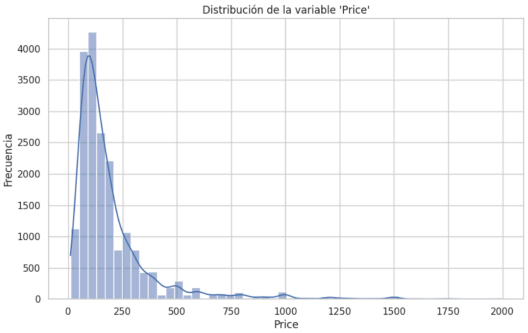
Notre objectif est donc d'analyser les données existantes afin de pouvoir prédire le prix des futurs logements en fonction de leurs caractéristiques. Nous identifierons les caractéristiques clés qui influencent le prix d'un logement. Nous veillerons à utiliser les outils vus en classe.

Description des données

Notre échantillon comprend 19 309 annonces et 29 variables avec certaines caractéristiques. Dans les 29 variables nous avons notamment la latitude et longitude géographique de la propriété soit la localisation, le nombre de pièces, le nombre de lits, s'il y a des frais de nettoyage ou pas, les avis des voyageurs, etc.

Partie 1 : Analyse Exploratoire des Données(EDA)

L'analyse exploratoire des données (EDA) permet comme son nom l'indique d'analyser et de visualiser les données pour en comprendre la structure, identifier les tendances et détecter les anomalies. Cela permet également de préparer les données pour le modèle. Nous avons commencé par visualiser les premières lignes, les informations générales de l'ensemble de données et à mettre en évidence les valeurs manquantes par colonne. Nous avons ensuite construit l'histogramme nous permettant d'obtenir la distribution de la variable prix puisque c'est celle ci qui nous intéresse en fonction de la fréquence.



Nous remarquons grâce à ce graphique que la plupart des logements ont des prix variants entre 50 et 250€.

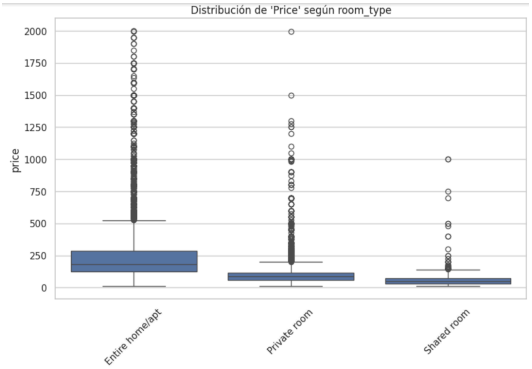
Nous allons à présent essayer de comprendre les relations entre les autres variables et notre variable cible. D'abord, nous commençons par l'encodage des variables catégoriques des colonnes catégorielles c'est-à-dire que nous avons transformé les

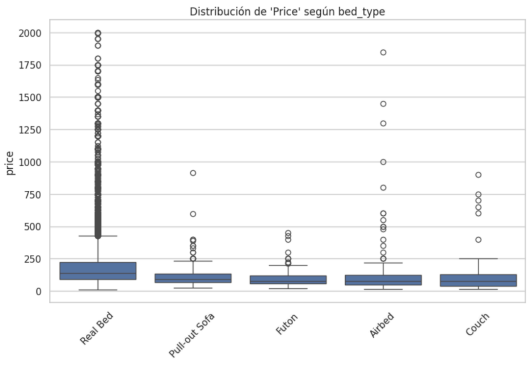
variables catégorielles en variables numériques à l'aide d'une colonne binaire pour chaque catégorie. Cela est nécessaire pour préparer les données pour nos analyses futures. Maintenant que nos données sont transformées, nous allons pouvoir identifier les corrélations qu'il peut y avoir avec les autres données. Cela nous permettra de comprendre quelles sont les caractéristiques qui ont une réelle influence sur le prix. Pour identifier les caractéristiques les plus influentes, nous avons affiché les 10 variables qui sont les plus corrélées avec le prix.

```
Variables les plus corrélées avec 'Price':
 price                          1.000000
accommodates                    0.442621
bedrooms                        0.437482
beds                            0.374791
bathrooms                       0.365740
city_SF                         0.112401
review_scores_rating            0.102133
neighbourhood_Capitol Hill      0.095454
city_DC                         0.072580
zipcode_94123                   0.072252
dtype: float64
```

Nous remarquons clairement que la capacité d'accueil (accommodates) est la variable la plus corrélée avec le prix avec un coefficient de 0.44. Ensuite, c'est le nombre de chambres avec 0.43. Et enfin, pour ne citer que les 3 premiers, le nombre de lits disponibles avec 0.36. Cela parait logique car ce sont des critères majeurs dans la tarification d'un logement. Grâce à ces informations, on sait que ces 3 variables seront probablement les plus utiles pour prédire le prix.

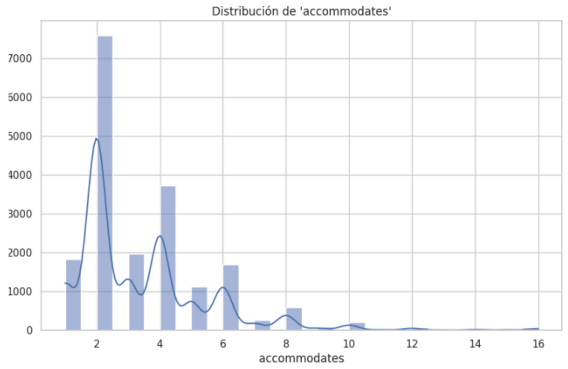
Ensuite, nous analysons les variables catégoriques qui sont les variables qui n'ont pas de données numériques comme le type de propriété par exemple. Il est important de les analyser car elles peuvent également avoir un très fort impact sur le prix. Nous les avons affichées sous forme de boxplot afin d'identifier rapidement la distribution des prix avec la médiane, la valeur minimale, la valeur maximale, etc. Cela permet aussi de mettre en évidence les outliers pour chaque catégorie.



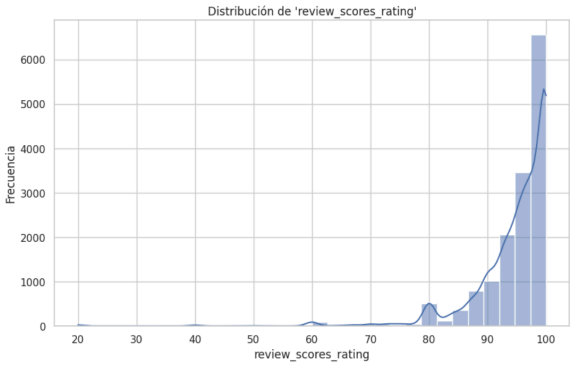


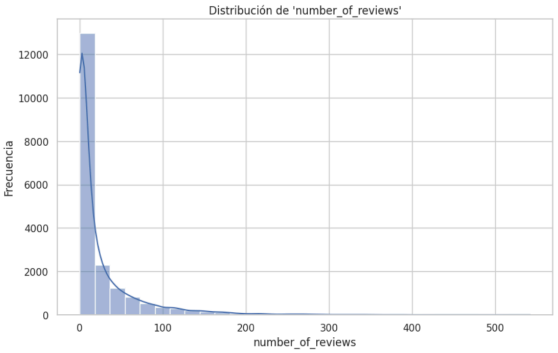
Nous remarquons que dans la catégorie type de logement, les prix sont nettement supérieurs lorsque la maison ou le logement est disponible dans son entièreté, même s'il y a beaucoup outliers (premier graphique). On constate également que lorsque le lit est un vrai lit et pas un canapé par exemple, le logement est plus cher (deuxième graphique).

Ensuite, nous avons réalisé des diagrammes pour avoir la distribution des variables numériques.



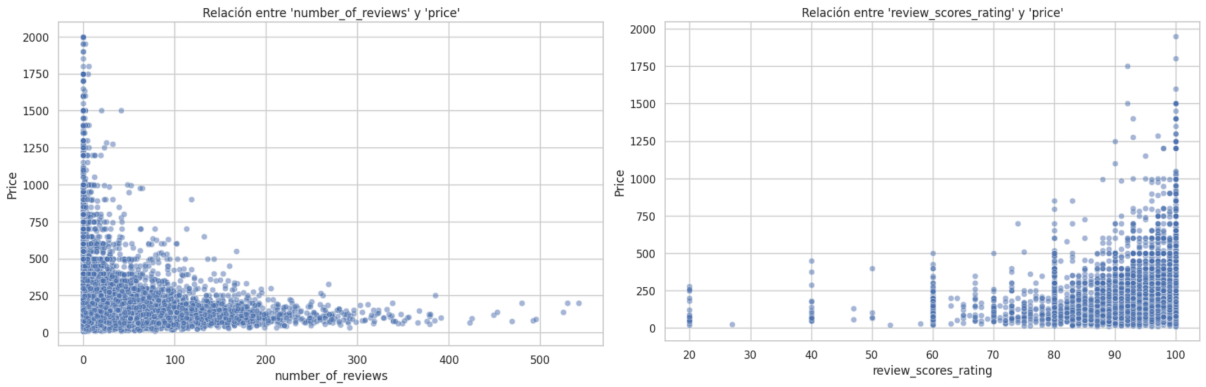
Lorsque nous nous intéressons à la capacité d'accueil, nous remarquons que la majorité des logements a une capacité d'accueil faible, 2 ou 4 personnes. Les logements ayant une capacité supérieure à 6 ou 8 personnes sont beaucoup moins fréquents. Les logements pouvant accueillir de grandes capacités (8 personnes ou plus) sont rares et pourraient être plus chers, influençant directement le prix moyen des logements. Nous pourrions également nous dire que les capacités très élevées (16 personnes) sont inhabituelles et peuvent correspondre à des logements spécifiques (grandes villas, propriétés pour événements) et auront donc un fort impact sur le prix aussi.





Sur le premier graphique, nous constatons aussi que le nombre de logements est croissant et que les notes supérieures à 8.5/10 sont nombreuses. On pourrait alors se dire que si les notes sont presque toutes très bonnes, elles vont avoir un faible impact sur le prix du logement. Ou au contraire, comme les notes sont très proches, elles vont avoir un grand impact car un rien fera la différence. Sur le deuxième graphique, une petite partie des logements concentre un grand nombre d'avis et une grande majorité des logements ont peu ou pas d'avis. Moins de logements ont un grand nombre d'avis, ce qui est logique car il faut du temps et un nombre élevé de réservations pour accumuler beaucoup d'avis. Les logements les plus populaires, les plus anciens ou ceux qui offrent une expérience exceptionnelle auront tendance à avoir plus d'avis, mais ils restent une minorité par rapport à la majorité des logements. Une grande partie des logements listés sur Airbnb n'ont reçu aucun avis. Cela peut indiquer qu'ils sont nouveaux ou qu'ils ne sont pas populaires par exemple.

A l'aide de graphiques de dispersion, nous allons à présent voir la relation entre les variables numériques et le prix.



Les deux graphiques ont des densités de points élevées. Cela signifie qu'il y a une concentration de données à cet endroit. D'après le premier graphique, on pourrait conclure que plus il y a d'avis sur les logements, plus le prix baisse. Et d'après le deuxième que plus la note du logement est élevée, plus le prix augmente.

Partie 2 : Machine learning

Nous avons commencé par importer les données. Ensuite, nous avons nettoyé les données c'est-à-dire que nous avons éliminé les colonnes qui n'étaient pas pertinentes et avons enlevé les valeurs nulles. Après nous avons séparé la variable cible (celle que nous cherchons à prédire) à savoir : le prix, et les variables prédictives soient les caractéristiques : nombre de chambres, localisation, etc. La variable cible est isolée pour que les modèles puissent apprendre à la prédire. Cela évite également d'utiliser la variable cible comme entrée, ce qui biaiserait l'entraînement du modèle. Nous avons continué en séparant les variables catégorielles et numériques. Cela permet de faire un prétraitement de chaque variable et de les rendre utilisables par les modèles. En ce qui concerne les variables numériques, elles sont directement utilisées pour les calculs mais peuvent nécessiter une mise à l'échelle (normalisation ou standardisation). Concernant les variables catégorielles, elles doivent être transformées en format numérique car les algorithmes de machine learning ne peuvent pas les traiter directement. Cela permet d'améliorer la précision du modèle pour éviter les interprétations erronées des données, et permet d'appliquer des techniques adaptées à chaque type de donnée.

Pour ce faire, nous avons réalisé notre pipeline en prétraitant les données (codage et standardisation). Nous avons ensuite divisé nos données en 3 afin de pouvoir utiliser 3 méthodes différentes.

D'abord, nous avons utilisé le modèle Random Forest qui est un algorithme d'apprentissage supervisé basé sur des ensembles d'arbres de décision. Chaque arbre est entraîné sur un échantillon aléatoire des données d'entraînement. Ensuite, nous avons divisé les données en formation et test. Nous avons donc un ensemble d'entrainement (70-80%) pour construire le modèle et un ensemble de tests (20-30%) pour évaluer les performances sur des données inconnues. Cela évite le surapprentissage et fournit une estimation réaliste de la capacité du modèle à généraliser. Nous avons effectué la validation croisée en utilisant 5 plis avec parallélisation. Cela permet d'évaluer la robustesse d'un modèle en divisant les données en plusieurs sous-ensembles. Ici, avec 5 plis, les données

sont divisées en 5 parties, le modèle est entraîné sur 4 plis et testé sur le 5e. Ce processus est répété 5 fois, en utilisant un pli différent pour tester à chaque fois. Cette méthode permet de fournir une évaluation plus stable de la performance et de réduire l'impact d'une division aléatoire des données sur les résultats. Nous avons ensuite affiché le coefficient R² moyen ainsi que le RMSE.

```
Evaluación del modelo RandomForest:
R2 Score: 0.4345
RMSE: 153.26
```

Le R² (le coefficient de détermination) est égal à 0.43, cette valeur indique que 43,45% de la variance totale des prix des logements est expliquée par le modèle. On peut dire que le modèle est correct mais encore améliorable. Le RMSE représente l'erreur moyenne quadratique entre les valeurs réelles et prédites. Cela signifie que les prédictions du modèle s'écartent en moyenne de 153.26 unités de la vraie valeur du prix. Ce qui n'est pas idéal étant donné que les prix varient principalement entre 50 et 250€.

Nous avons également utilisé le Support Vector Regression (SVR) qui est une méthode de régression. Cette méthode permet de prédire une valeur continue en minimisant l'erreur. C'est une méthode qui permet d'optimiser les erreurs de régression tout en limitant la complexité du modèle.

Pour finir, nous avons utilisé le PCA (Principal Component Analysis). Le PCA est une méthode non supervisée qui permet de réduire la dimension, le bruit et de faciliter la visualisation en améliorant les performances du modèle. Il existe 2 méthodes : la sélection et l'extraction. La sélection permet de sélectionner des données par la variance, la corrélation ou l'importance. L'extraction permet de créer de nouvelles caractéristiques à partir d'anciennes grâce aux mathématiques. D'abord, il y a la standardisation, puis le calcul de la matrice de covariance. Ensuite, il y a la décomposition en valeur propre et enfin la projection des données.

Nous avons choisi le random forest car c'était le modèle le plus performant. Une fois le modèle validé, nous avons pu former le modèle final en le reconstruisant avec l'intégralité de l'ensemble d'entraînement. Nous avons effectué les prédictions et l'évaluation, c'est-à-dire que nous avons appliqué le modèle final sur l'ensemble de test pour prédire les prix et nous avons comparé ces prédictions avec les vraies valeurs pour évaluer la performance de test. Si l'on garde un aspect critique random forest ne devrait pas être l'outil le plus performant, peut être que le PCA et le choix des colonnes à garder est à revoir.

Conclusion

Lors de ce projet, nous avons pu mettre en application toutes les techniques vues en cours. Nous avons réussi à simplifier les données de l'échantillon et à rendre cet échantillon exploitable par les différents modèles : ECA, Random Forest, SVR, PCA… A l'aide des différents graphiques et données d'évaluation nous avons pu identifier quelles étaient les variables corrélées avec notre valeur cible : le prix. Nous avons à présent un modèle qui est capable de prédire le prix de nouveaux logements (avec une marge d'erreur assez élevée certes).

Références

- Image logo airbnb : https://www.google.com/search?q=airbnb&sca_esv=41165fec71bc817d&rlz=1C1ASVC_frFR918FR918&biw=767&bih=730&sxsrf=ADLYWIKuapWpk19WyAbfURXF6v78s3Zxhg:1732709497787&source=lnms&fbs=AEQNm0CbCVgAZ5mWEJDg6aoPVcBg-SdB_VhkTgIuMkvHavRAAqJIdK1d4IRkOOfjc8W6G2-SHa85SoGxCowoDAcZOVZC68ahKX45xe2rydIORCVv2PAs1hiz6ZvcAKhtNUMgPdjzbvULzKJPMwD9zMBDiEl52FoDqhbBy7BGIFSDUd_jYhKqhZCZpiCWW18XVIOE0LajBZc26WObl4RWYC7BBHiXeOTuIA&sa=X&ved=2ahUKEwj5_PTwvfyJAxUIGLkGHQVsMfYQ0pQJegQIBhAD#vhid=uuVr-sdgEuMFnM&vssid=_egxHZ9uVNLre5OUP-PbwmQY_98

- Cours Ciencia de los Datos

- Définition ECA : https://www.ibm.com/fr-fr/topics/exploratory-data-analysis#:~:text=Qu'est%2Dce%20que%20l'EDA%20%3F,m%C3%A9thodes%20de%20visualisation%20des%20donn%C3%A9es.
- Définition SVR : https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2