Immobiliers Silicon Valley

SOMMAIRE

1) Contexte

2) EDA

3) MODEL

4) Axes d’ameliorations

Contexte

Votre demande :

Créer un modèle prédictif pour prédire la valeur des logements en Californie

Les outils apportés :

Une base de données qui contient les prix médians des logements pour les districts de Californie issus du recensement de 1990

Rendu :

03/02/23

Contexte

Planning previsionnel :

Janvier :

12 : EDA + Baseline / 13 : Iteration , cross validation, nettoyage des donnés, veille knn / 25 : scaling

Fevrier :

1 : Pipelines pour industrialiser le processus (automatiser) / 2 et 3 : Model hyperparametre

Planning réalisé :

Janvier :

12 : Creation repository local et remote + Création planning + veille : EDA + Debut EDA + creation Figma / 13 : EDA + veille alias + creation alias / 15 : EDA / 17 : EDA + Planning update / 25 : EDA + veilles : Data preprocessing + Debut model / 26 : Baseline / 31 : Iterations + EDA cleaning

Fevrier :

1 : Gros probleme avec Git jusqu'a 14h10 + refaire rajout EDA + nouvelles iterations + check model / 2 : Nouvelles iterations + Model Hold Out, Regression Lineai / 3 : Check EDA et model + presentation

Rendu :

03/02/23

EDA

1) Imports librairies + train_data

	Unnamed: 0	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proxin
0	2072	-119.84	36.77	6.0	1853.0	473.0	1397.0	417.0	1.4817	72000.0	INLA
1	10600	-117.80	33.68	8.0	2032.0	349.0	862.0	340.0	6.9133	274100.0	<1H OCE
2	2494	-120.19	36.60	25.0	875.0	214.0	931.0	214.0	1.5536	58300.0	INLA
3	4284	-118.32	34.10	31.0	622.0	229.0	597.0	227.0	1.5284	200000.0	<1H OCE
4	16541	-121.23	37.79	21.0	1922.0	373.0	1130.0	372.0	4.0815	117900.0	INLA
5	8781	-118.32	33.79	32.0	2381.0	467.0	1264.0	488.0	4.1477	315100.0	<1H OCE
6	5438	-118.43	34.01	31.0	2526.0	528.0	1046.0	504.0	4.7009	500001.0	<1H OCE
7	14856	-117.07	32.64	32.0	5135.0	1025.0	2152.0	944.0	4.1325	172800.0	NEAR OCE
8	19956	-119.33	36.22	9.0	3748.0	644.0	1955.0	620.0	4.2011	108100.0	INLA
9	17175	-122.47	37.50	18.0	2297.0	416.0	1086.0	381.0	4.8750	334600.0	NEAR OCE

EDA

2) Observation du data set

dimension

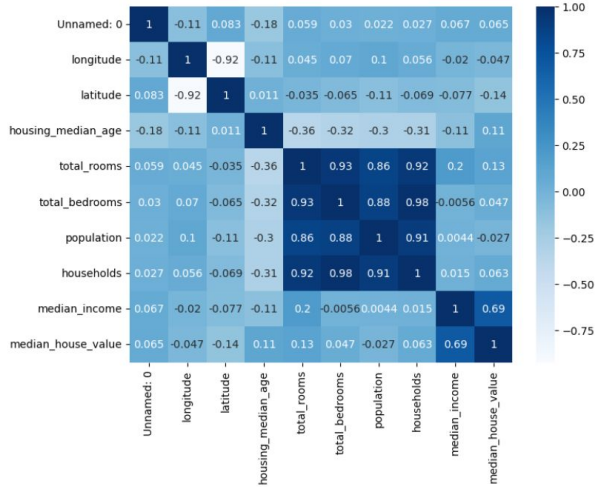
valeur min

corr = total_rooms, households, population, total_bedrooms, population

types

moyenne

NaN



EDA

3) Nettoyage et transformation

Rename population total_residents

OneHotEncoder on ocean_proximity

Remove columns: unnamed: 0, latitude and longitude

ocean_proximity	ocean_proximity_<1H OCEAN	ocean_proximity_INLAND	ocean_proximity_ISLAND	ocean_proximity_NEAR BAY	ocean_proximity_NEAR OCEAN
INLAND	0.0	1.0	0.0	0.0	0.0
<1H OCEAN	1.0	0.0	0.0	0.0	0.0
INLAND	0.0	1.0	0.0	0.0	0.0
<1H OCEAN	1.0	0.0	0.0	0.0	0.0
INLAND	0.0	1.0	0.0	0.0	0.0
...	...	...	...	...	...
INLAND	0.0	1.0	0.0	0.0	0.0
NEAR BAY	0.0	0.0	0.0	1.0	0.0
INLAND	0.0	1.0	0.0	0.0	0.0
<1H OCEAN	1.0	0.0	0.0	0.0	0.0
NEAR OCEAN	0.0	0.0	0.0	0.0	1.0

EDA

3) Nettoyage et transformation

Null handling

mean
median
delete

Exports dataframe

total_bedrooms
NaN
NaN
NaN
NaN
NaN
...
NaN

Outliers

IQR method to find outliers

Global visualization of outliers

MODEL

1) Imports librairies + iterations

MODEL

2) Models utilisés

- Dummy
- Hold Out
- Regression Lineaire
- Cross validation

MODEL

2) Iteration utilisés

- immo_base
- df_no_outliers

1

- df_immobilier_no_null
- df_immobilier_null_mean
- df_immobilier_null_median

Amélioration

- Techniques :
  - Connaissances et utilisation des modèles : RandomForest, KNN,..
  - .gitignor
  - Pickle

- Gestion projet :
  - Mauvaise compréhension des consignes (EDA 2/3j)
  - Mauvaise compréhension du référentiel (Objectif 1?)

- Bilan :
  - Reprendre de mon côté le projet pour maîtriser les techniques utiliser

Feed back

- Pour plus de performance, une base de données plus récentes

Avez-vous des questions ?

Référence

- notebook evaluation model
- notebook diabetes
- https://medium.com/analytics-vidhya/different-type-of-feature-engineering-encoding-techniques-for-categorical-variable-encoding-214363a016fb
- https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing
- https://towardsdatascience.com/preprocessing-with-sklearn-a-complete-and-comprehensive-guide-670cb98fcfb9
- Aides camarades

Merci