Predict Medical Desert

IRONHACK Final Project

Ludivine Lacour

CONTEXT & PROBLEM DEFINITION

Medical Desert in France

Some people have to drive many kilometers to find a doctor which can be hard for some of them having a limited way of commute (old people, people without any driving license).

APL indicator: Potential nb of consultations/resident/year (around 20 min drive)

Can we predict the medical desert of an area?
What factors would impact the lack of doctors in an area?

PROCESS

01

Data collection and cleaning

Find relevant data to list of assumptions.

Merging data sources.

Creation of calculated columns.

02

Exploratory

Exploration of target.

Correlation of data.

Linearity relationship between target and features.

03

Classification

Compare classification models using pycaret.

Logistic Regression and Gradient Boosting.

04

Feature Importances

Feature Engineering using ANOVA F measure and SFS.

Decision Tree and Gradient Boosting, feature importance.

Data Collection

What could be the factors of medical desert?

01 DATA COLLECTION & CLEANING

100K

General practitioners in France

~60 000

would be free to choose where they want to practice

Assumptions on factors impacting medical desert in a city:

- **Population / area density**
- **Population growth**
- **Population average age**
- **Birth rate**
- **Socio-Professional Category**
- **Level of poverty**
- **Unemployment rate**
- **Number of medical infrastructures**
- **Level of medical education**
- **Level of city amenities / investment in city amenities**
- **Expense in healthcare (per resident in a city)**
- **Average temperature**

01 DATA COLLECTION & CLEANING

Data sources:

- **APL indicator**: data.drees.sante.gouv.fr
- **Data for calculated metrics**: INSEE.fr (several data sources)

	REG	DEP	DEPCOM	DCIRIS	AN	TYPEQU	NB_EQUIP
0	84	1	1001	01001	2018	A401	2
							4
							1
							1
							1

	CODGEO	P16_POP	P16_POP0014	P16_POP1529	P16_POP3044	P16_POP4
0	1001	767.0	161.000000	102.000000	132.000000	189.000
						000
						528
						691
						000

	CODGEO	LIBGEO	REG	DEP	P16_POP	P11_POP	SUPERF	NAIS1116	DECE
0	01001	L'Abergement-Clémenciat	84	01	767	780	15.95	41	
1	01002	L'Abergement-de-Varey	84	01	243	234	9.15	21	
2	01004	Ambérieu-en-Bugey	84	01	14081	13839	24.60	1114	
3	01005	Ambérieux-en-Dombes	84	01	1671	1600	15.92	101	
4	01006	Ambléon	84	01	110	112	5.88	9	

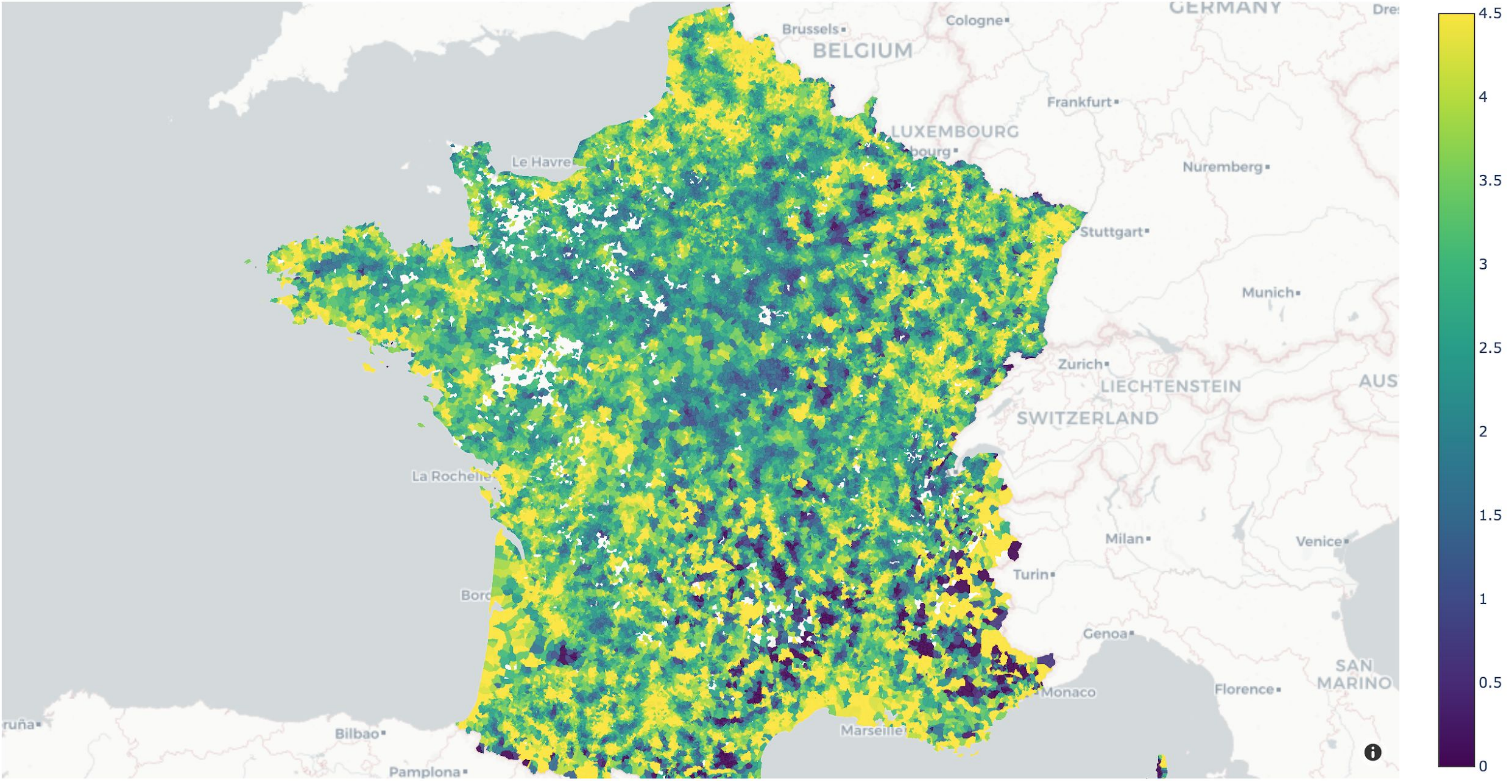
median_living_standard	healthcare_education_establishments	density_area	annual_pop_growth	unemplo
22679.000000	0	48.087774	-0.335578	
24382.083333	0	26.557377	0.757662	
19721.000000	0	572.398374	0.347315	
23378.000000	0	104.962312	0.872154	
21660.000000	0	18.707483	-0.359722	
22146.451613	0	80.000000	2.562896	
24893.809524	0	143.678161	0.432215	
23088.000000	0	48.414986	-0.177621	
22880.555556	0	38.414217	1.742295	

34989 rows (cities)
21 features (factor assumptions)

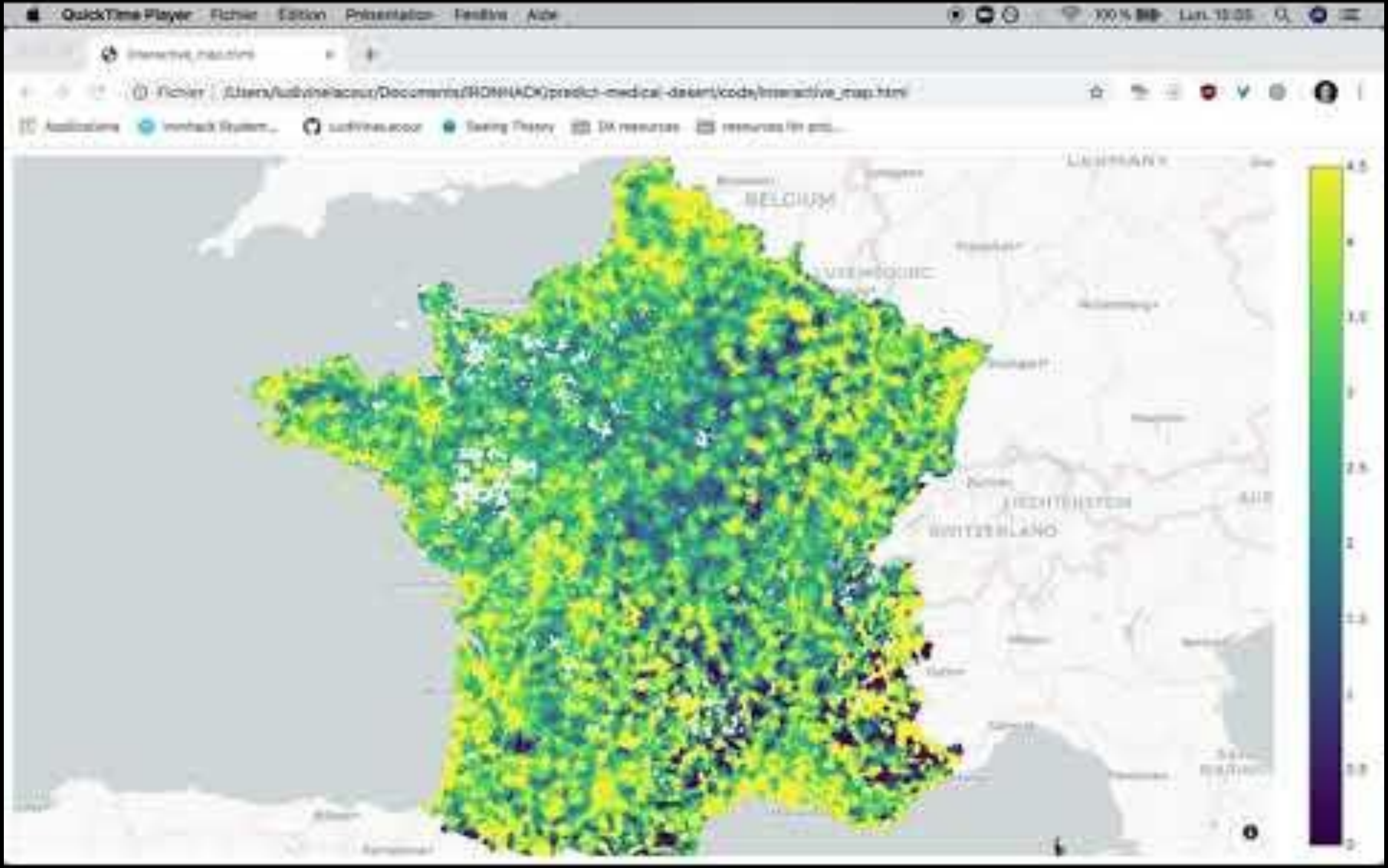
Exploratory analysis

What can we learn from our data?

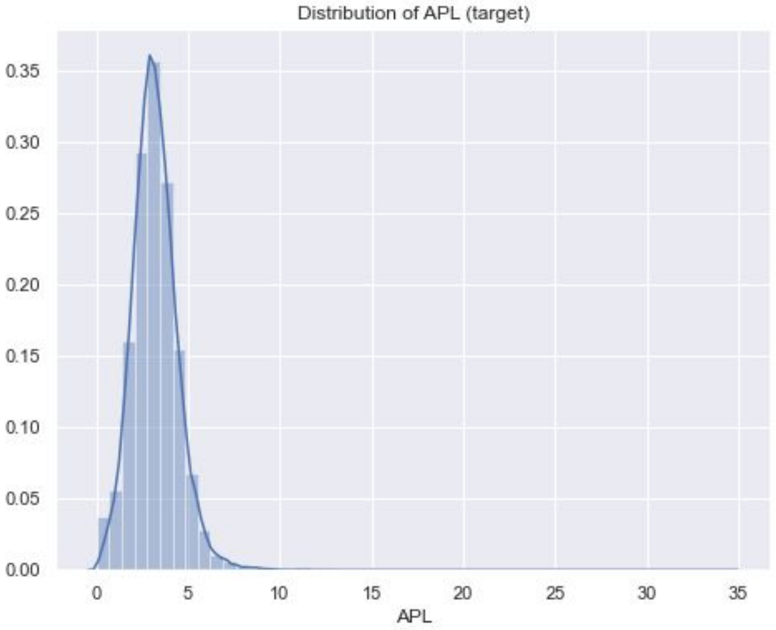
02 EXPLORATORY

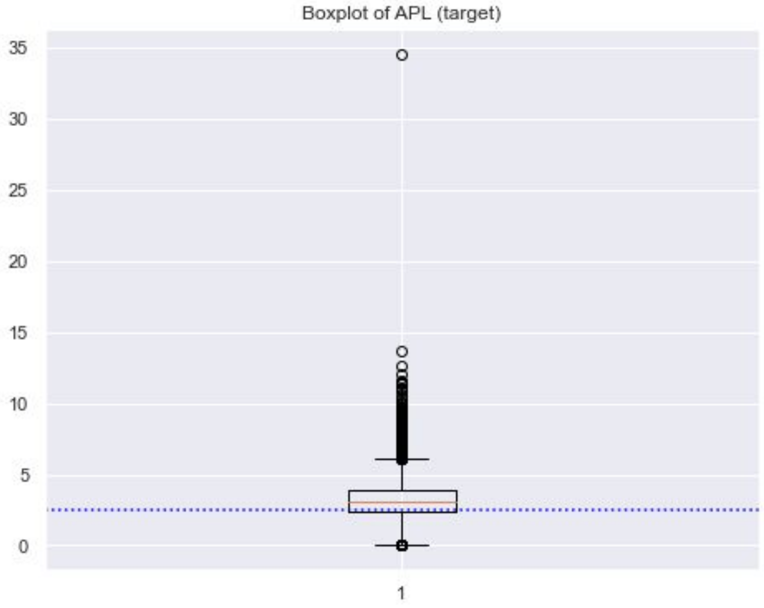


https://www.youtube.com/watch?v=WIgnRYnvtyk



02 EXPLORATORY

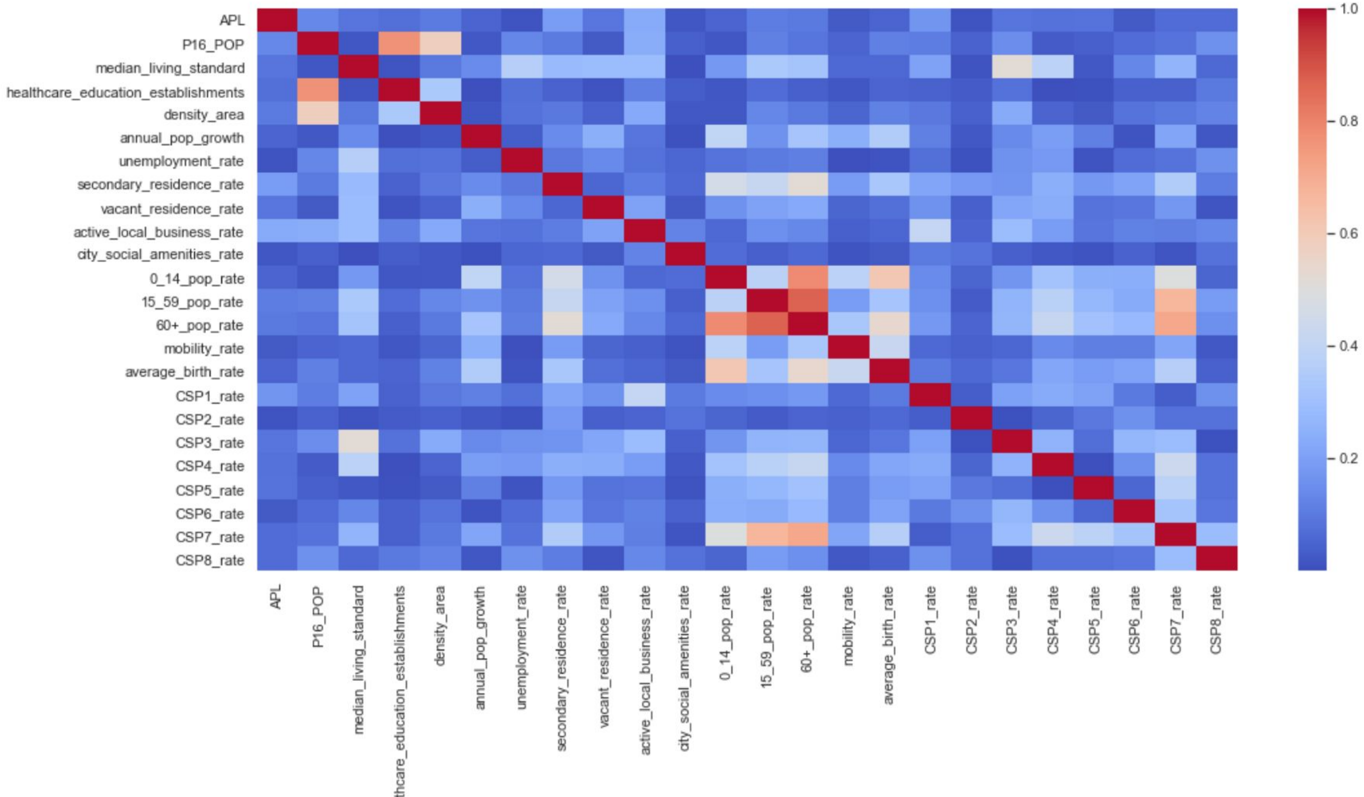




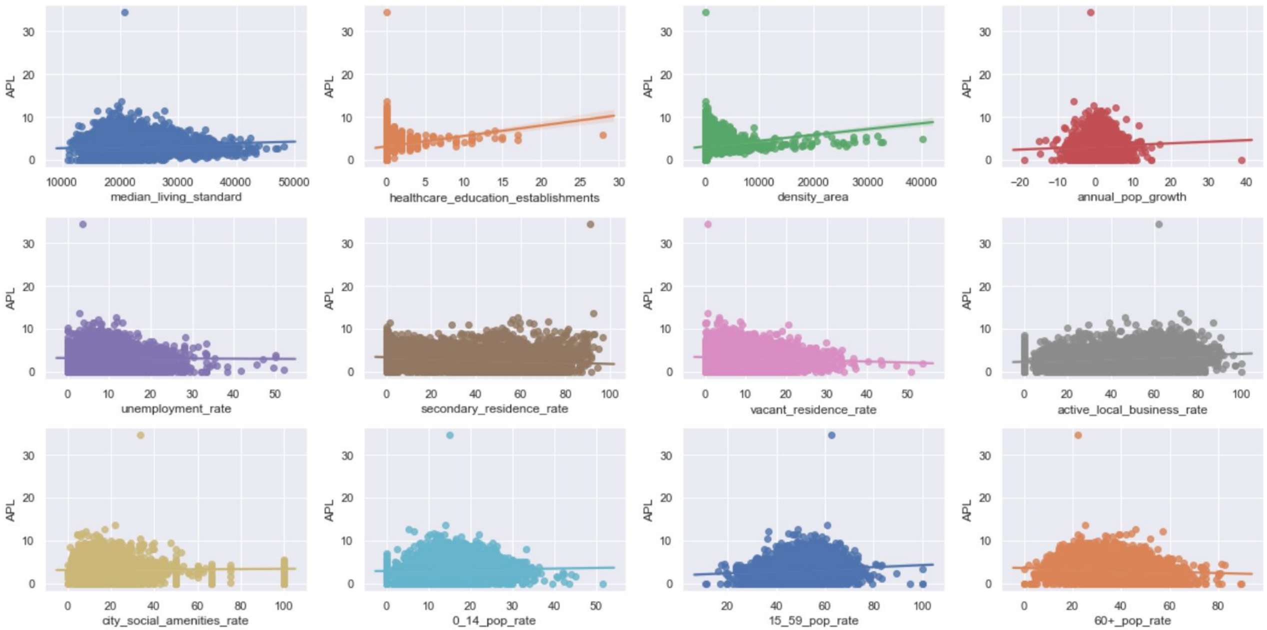
⇒ target is kinda normally distributed, we can disparity between city having huge available consultations and some having no consultation.

02 EXPLORATORY

Correlation of features



02 EXPLORATORY



⇒ Absence of linearity relationship between target and features

Classification

Can we predict a medical desert?

03 CLASSIFICATION

Split the data into 3 categories:

No medical desert	Potential medical desert	Medical desert
18751 cities	10719 cities	5519 cities

⇒ Build an algorithm that predicts in which category the city belongs given the 21 features.

03 CLASSIFICATION

Model used: Logistic Regression model

(Train sample = 11589 obs., Test sample = 4968 obs., 21 features)

medical_desert = 0	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
const	-0.5309	1.5190	-0.3495	0.7267	-3.5081	2.4462
median_living_standard	6.2835	1.1208	5.6060	0.0000	4.0867	8.4803
healthcare_education_establishments	-0.3214	0.1505	-2.1353	0.0327	-0.6165	-0.0264
density_area	-0.5984	1.4274	-0.4193	0.6750	-3.3960	2.1991
annual_pop_growth	-3.0637	1.5685	-1.9533	0.0508	-6.1379	0.0105
unemployment_rate	0.6596	0.7516	0.8776	0.3801	-0.8135	2.1327
secondary_residence_rate	1.7063	0.2282	7.4769	0.0000	1.2590	2.1535
vacant_residence_rate	2.5734	0.5850	4.3992	0.0000	1.4269	3.7199
active_local_business_rate	-1.5820	0.1834	-8.6265	0.0000	-1.9415	-1.2226
city_social_amenities_rate	-1.6776	0.3391	-4.9472	0.0000	-2.3422	-1.0130
0_14_pop_rate	4.1817	0.7781	5.3743	0.0000	2.6567	5.7067
15_59_pop_rate	-0.6419	0.5941	-1.0805	0.2799	-1.8062	0.5225
mobility_rate	3.3470	3.9961	0.8375	0.4023	-4.4853	11.1792
average_birth_rate	2.5867	7.8517	0.3294	0.7418	-12.8024	17.9758
CSP1_rate	1.4229	1.5841	0.8982	0.3691	-1.6820	4.5277
CSP2_rate	-1.1742	1.6037	-0.7322	0.4641	-4.3174	1.9690
CSP3_rate	-0.9321	1.5699	-0.5937	0.5527	-4.0091	2.1448
CSP4_rate	-0.6425	1.5216	-0.4222	0.6729	-3.6248	2.3399
CSP5_rate	-1.1115	1.5154	-0.7335	0.4633	-4.0817	1.8586

⇒ All features are not relevant but even when selecting the best features, the model performance doesn't improve.

All features (21)

```
=====================================================
None of feature selection
Accuracy score for train sample: 0.48347570972473897
Accuracy score for test sample: 0.47061191626409016
Classification report:
              precision    recall  f1-score   support

           0       0.47      0.54      0.50      1673
           1       0.40      0.35      0.37      1657
           2       0.53      0.52      0.53      1638

    accuracy                           0.47      4968
   macro avg       0.47      0.47      0.47      4968
weighted avg       0.47      0.47      0.47      4968
```

Sequential Forward Selection features (4)

```
=====================================================
SFS_4 feature selection
Accuracy score for train sample: 0.4717404435240314
Accuracy score for test sample: 0.46557971014492755
Classification report:
              precision    recall  f1-score   support

           0       0.47      0.55      0.50      1673
           1       0.40      0.33      0.36      1657
           2       0.52      0.51      0.52      1638

    accuracy                           0.47      4968
   macro avg       0.46      0.47      0.46      4968
weighted avg       0.46      0.47      0.46      4968
```

03 CLASSIFICATION

Pycaret Library comparison models

All features (21 features)

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
0	Extreme Gradient Boosting	0.5184	0.0	0.5184	0.5146	0.5148	0.2776
1	Gradient Boosting Classifier	0.5160	0.0	0.5160	0.5121	0.5122	0.2740
2	Light Gradient Boosting Machine	0.5155	0.0	0.5155	0.5120	0.5122	0.2732
3	Ada Boost Classifier	0.5070	0.0	0.5070	0.5032	0.5040	0.2606
4	Extra Trees Classifier	0.5065	0.0	0.5065	0.5025	0.5018	0.2598
5	Random Forest Classifier	0.4849	0.0	0.4849	0.4832	0.4795	0.2273
6	Linear Discriminant Analysis	0.4829	0.0	0.4829	0.4775	0.4768	0.2243
7	Ridge Classifier	0.4818	0.0	0.4818	0.4744	0.4714	0.2228
8	Logistic Regression	0.4812	0.0	0.4812	0.4752	0.4740	0.2219
9	SVM - Linear Kernel	0.4673	0.0	0.4673	0.4606	0.4250	0.2009
10	K Neighbors Classifier	0.4364	0.0	0.4364	0.4389	0.4359	0.1546
11	Naive Bayes	0.4289	0.0	0.4289	0.5245	0.3767	0.1433
12	Decision Tree Classifier	0.4205	0.0	0.4205	0.4326	0.3955	0.1307
13	Quadratic Discriminant Analysis	0.4196	0.0	0.4196	0.5197	0.3547	0.1294

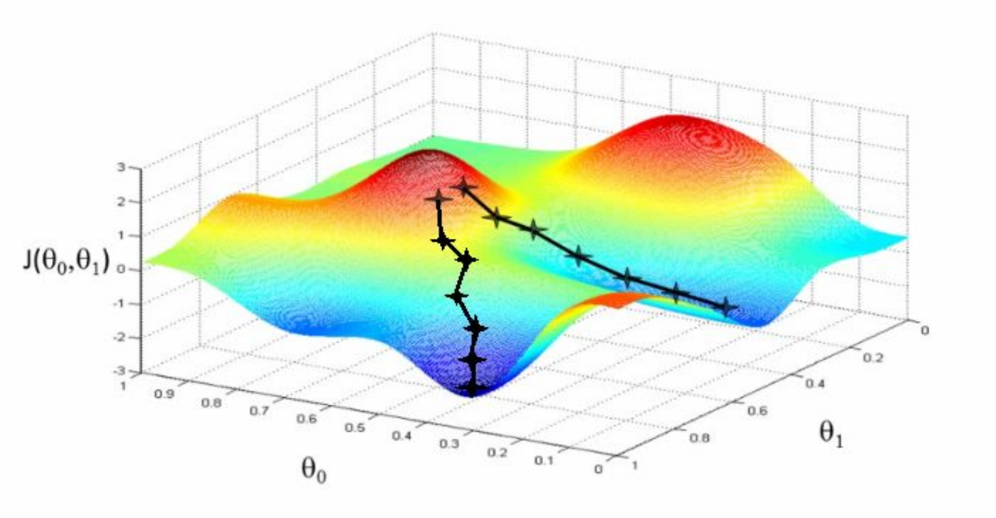
"Best" features (4 features)

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
0	Extreme Gradient Boosting	0.4767	0.0	0.4767	0.4756	0.4757	0.2151
1	Gradient Boosting Classifier	0.4750	0.0	0.4750	0.4736	0.4738	0.2125
2	Light Gradient Boosting Machine	0.4700	0.0	0.4700	0.4675	0.4681	0.2050
3	Ada Boost Classifier	0.4690	0.0	0.4690	0.4676	0.4677	0.2035
4	Ridge Classifier	0.4649	0.0	0.4649	0.4592	0.4539	0.1974
5	Linear Discriminant Analysis	0.4637	0.0	0.4637	0.4600	0.4570	0.1956
6	Logistic Regression	0.4635	0.0	0.4635	0.4579	0.4545	0.1953
7	Quadratic Discriminant Analysis	0.4609	0.0	0.4609	0.4662	0.4540	0.1913
8	SVM - Linear Kernel	0.4541	0.0	0.4541	0.4277	0.3861	0.1811
9	Naive Bayes	0.4483	0.0	0.4483	0.4522	0.4364	0.1724
10	Extra Trees Classifier	0.4420	0.0	0.4420	0.4407	0.4410	0.1630
11	Random Forest Classifier	0.4385	0.0	0.4385	0.4379	0.4361	0.1578
12	K Neighbors Classifier	0.4175	0.0	0.4175	0.4197	0.4140	0.1262
13	Decision Tree Classifier	0.3949	0.0	0.3949	0.4032	0.3730	0.0924

⇒ Gradient Boosting could be a model that works to predict medical desert but performance should be improved.

03 CLASSIFICATION

Model used: Gradient Boosting Classifier



Accuracy of the model
56%

⇒ Gradient Boosting minimize the errors between each direction using the Gradient Descent Algorithm.

03 CLASSIFICATION

Running Gradient Boosting model with Stratified cross-validation

```
# Using StratifiedKFold for cross-validation
accuracies_train=[]
accuracies_test=[]
skf = StratifiedKFold(n_splits=10, random_state=8, shuffle=True)
gradient = GradientBoostingClassifier(random_state=8)

for train_idx, test_idx in skf.split(X_res,y_res):
    gradient = gradient.fit(X.iloc[train_idx,:],y[train_idx])
    accuracies_train.append(accuracy_score(y[train_idx],gradient.predict(X.iloc[train_idx,:])))
    accuracies_test.append(accuracy_score(y[test_idx],gradient.predict(X.iloc[test_idx,:])))

print("Average accuracy for train samples:",np.mean(accuracies_train))
print("Average accuracy for test samples:",np.mean(accuracies_test))
```

```
Average accuracy for train samples: 0.611463417694258
Average accuracy for test samples: 0.5614536538377337
```

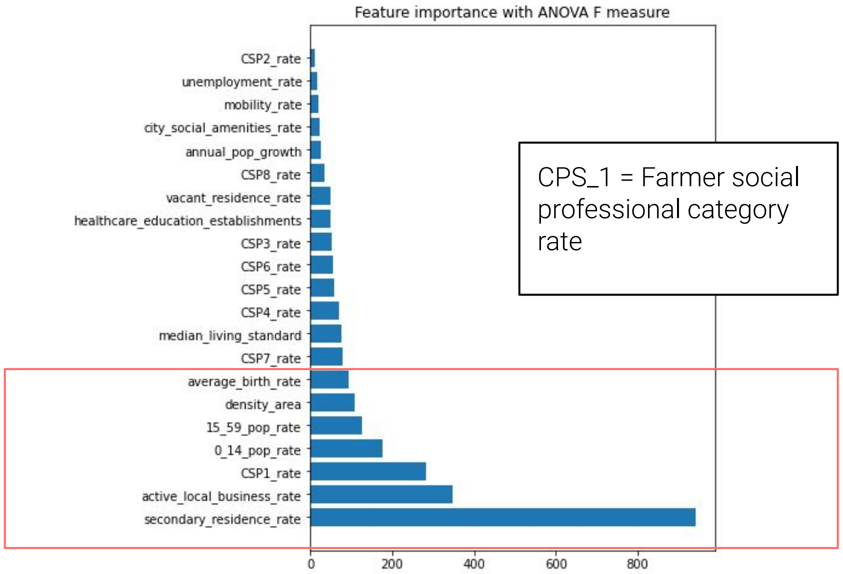
⇒ We should improve the model with hyperparameter tuning or by using other pre-processing methods.

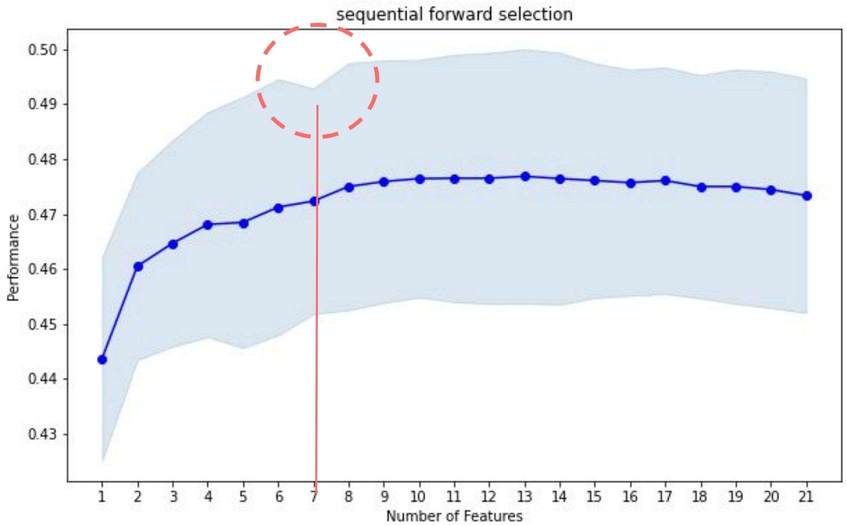
Feature importance

What factors increase the chance of having a medical desert?

04 FEATURE IMPORTANCE

Feature Engineering (ANOVA & Sequential Forward Selection)

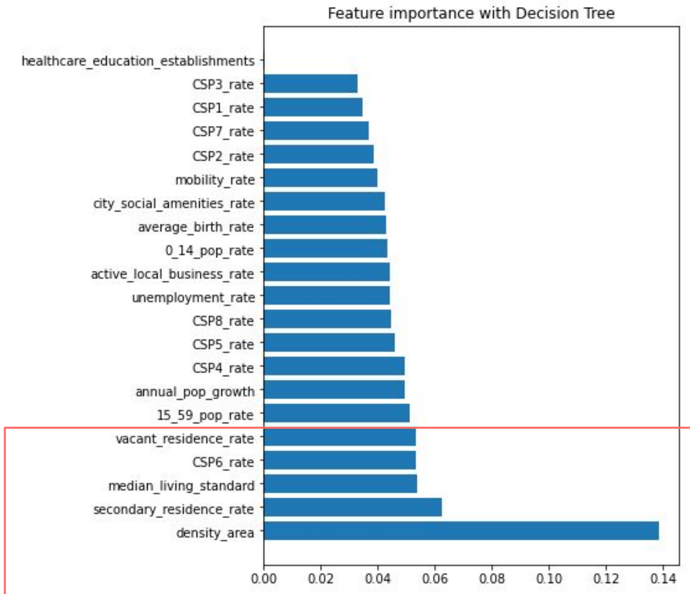




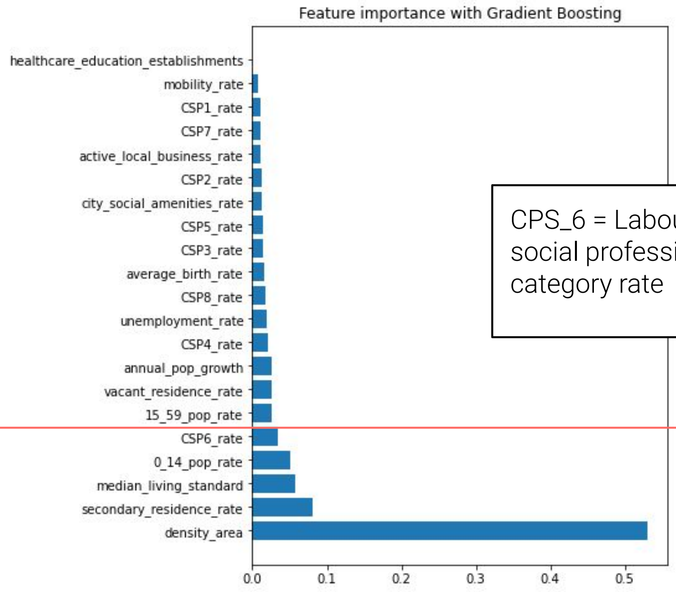
⇒ First list of important features and 7 features could be the right number of feature

04 FEATURE IMPORTANCE

Decision Tree



Gradient Boosting Tree



CPS_6 = Labourer social professional category rate

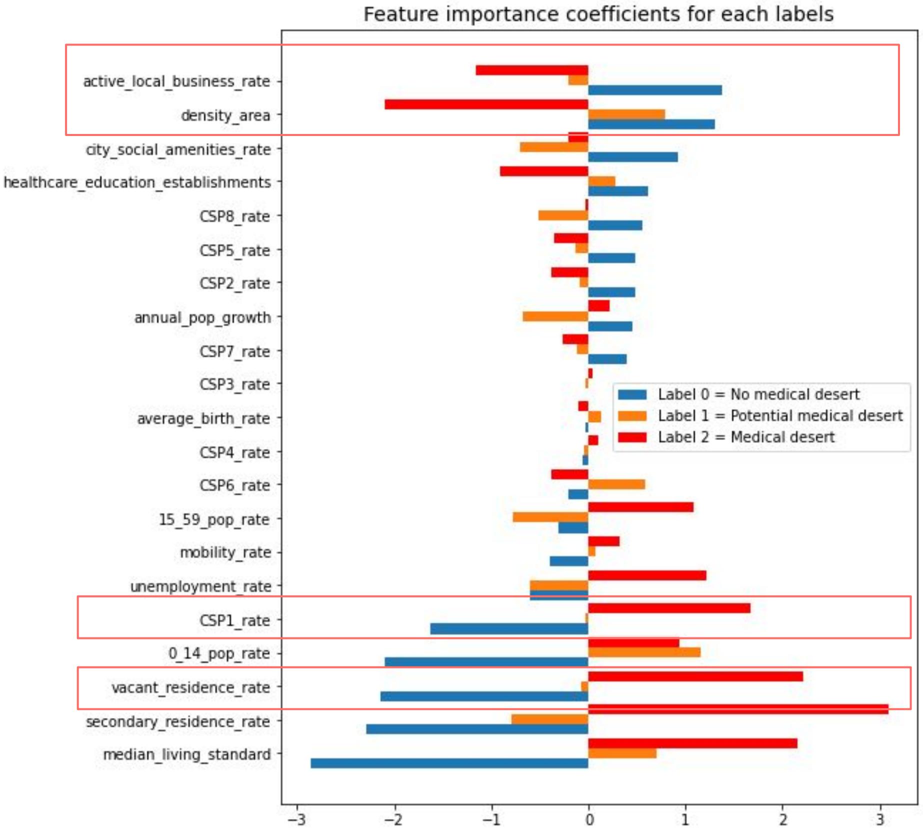
⇒ Important feature are almost the same

04 FEATURE IMPORTANCE

⇒ Rural environment is definitely a place where doctors are missing:

- More CSP1 (farmer) and many vacant residences increase chances of being in a medical desert

- Having many local business and services and more resident per km2 decrease chances of medical desert

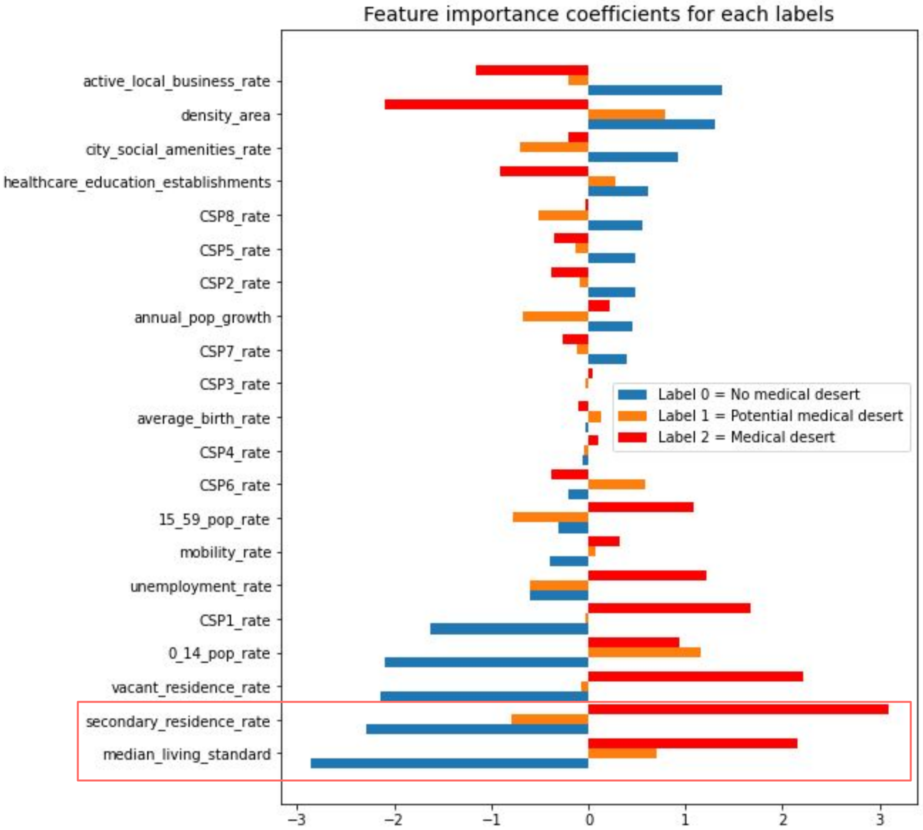


04 FEATURE IMPORTANCE

⇒ High standard of living would bring more to medical desert:

- Population with high salary (median living standard) would have less difficulties accessing doctors so less concerns about living near doctors

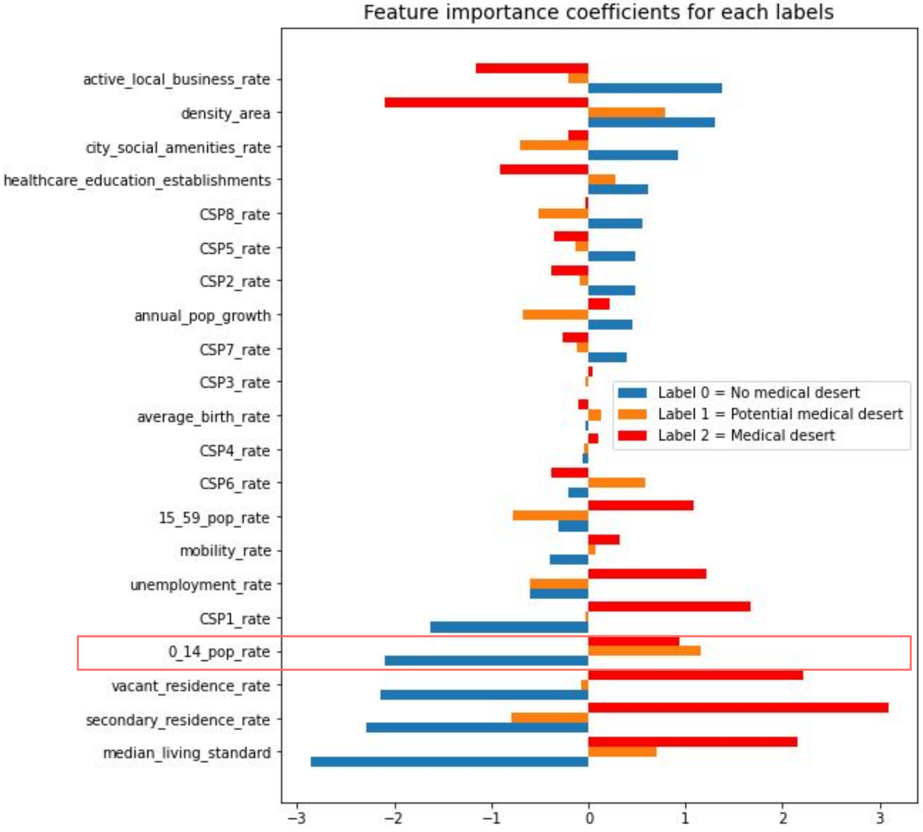
- There is less need of being seen in area of secondary resident (occupation rate is lower)



04 FEATURE IMPORTANCE

⇒ Cities with an higher ratio of children would lack doctors:

- Knowing children have more need of consultation, the demand is higher



CONCLUSION

Medical desert prediction can be improved

But Gradient Boosting Classification could be the right model.

Interesting insights on factors affecting medical desert

Manage to know more or less on what factors a city can play to influence medical desert.

POSSIBLE IMPROVEMENTS:

- **Improve Gradient Boosting algorithm:** work on hyper tuning parameters or other preprocessing methods
- **Iteration of class threshold**
- **PCA:** reduce the nb of columns now I have vision on feature importances
- **Bin of continuous variables:** to avoid issue on classification models
- **Create clusters of medical deserts**: see if we can group different type of medical desert