



Participez à une compétition Kaggle GoDaddy – Microbusiness Density Forecasting

Parcours IML - OpenClassRooms - CentraleSupélec
Tatiana Martinez

Sommaire

CONTEXTE

Objectif
Données

01

NETTOYAGE DES DONNÉES

Les étapes

02

MANIPULATIONS

Les étapes
Visualisations

03

04

FEATURE ENGINEERING

Création de nouvelles variables

05

EXPLORATION DES DONNÉES

Visualisations

06

MÉTHODES et OBJECTIFS

Preprocessing
Stratification
Models et Scores
Sample Submission

Perspectives et Questions



01

CONTEXTE

- Kaggle : une plateforme qui organise des compétitions en data science
- Compétition organisée par Go Daddy, une entreprise états-unienne de services pour les entrepreneurs
 - Les microentreprises aux USA n'apparaissent pas toujours dans les sources économiques classiques.
 - Les études Kaggle devront permettre aux décideurs de pouvoir les étudier et mieux comprendre les facteurs associées à celles-ci.

OBJECTIF

Prédire la densité des microentreprises aux USA par comté pour les périodes 01-11-2022 au 01-06-2023 grâce à des données de 01-08-2019 au 01-10-2022

Données

4 fichiers :

- Fichier train.scv qui fournit **l'activité mensuelle** des microentreprises par comté
- Fichier test.csv qui fournit les **références des microentreprises sur une période**
- Fichier census_starter.csv qui fournit des informations de **recensement par comté** avec 2 ans de "retard" par rapport aux données de densité des microentreprises
- Fichier sample_submission qui fournit des **valeurs exemples d'activités des microentreprises au format attendu pour la compétition**

Indicateur à prédire :

Density microbusiness par mois par comté

*Population de 18 ans et plus du comté

$$\frac{\text{Nombre brut de microentreprises actives dans le comté sur le mois}}{\text{Population de 18 ans et plus du comté}} \times 100$$

* les chiffres de population utilisés pour calculer la densité sont décalés de deux ans
Les chiffres de densité de 2021 sont calculés à partir des chiffres de population de 2019

Métrique à réduire :

Symmetric Mean Absolute Percentage Error

$$\text{SMAPE} = \frac{1}{n} \sum_{t=1}^n \frac{|F_t - A_t|}{(A_t + F_t)/2}$$






where A_t is the actual value and F_t is the forecast value.






SMAPE

Aperçu des données

Données de recensement par comté par an de 2017 à 2021

-  % des ménages ayant accès à Internet, haut débit
-  % de la population de + 25 ans avec Bac+4
-  % de la population né hors sol US
-  % des salariés employés dans l'industrie de l'information
-  Salaire médian des ménages

Données de densité des microentreprises

-  Par comté
-  Par mois du 01/08/2019 au 01/10/2022
-  Le nombre brut de microentreprises dans le comté

Variable à prédire

microbusiness_density

La densité des microentreprises, est obtenue en divisant la population de 18 ans et plus par le nombre brut de microentreprises actives dans une zone géographique, et en multipliant par 100. Les chiffres de la population accusent un décalage de deux ans en raison du rythme de mise à jour fourni par le U.S. Census Bureau, qui fournit chaque année les données démographiques sous-jacentes. Les chiffres de densité de 2021 sont calculés à partir des chiffres de population de 2019, etc.

KERNEL

Code(s) déposé(s) sur kaggle par un
candidat à la même compétition

<https://www.kaggle.com/code/titericz/better-xgb-baseline>

Commentaire Kaggle



tm.kaggle

Posted just now · Posted on Version 10 of 10

Hello, thank you for sharing. I've learned a lot with your notebook.

Concerning the part of outliers, may you explain the code (the rules) :

```
for i in range(37, 2, -1):
```

```
thr = 0.20*np.mean(var[:i])
```

```
difa = abs(var[i]-var[i-1])
```

```
if (difa>=thr):
```

```
var[:i] *= (var[i]/var[i-1])
```

```
outliers.append(o)
```

```
cnt+=1
```

```
var[0] = var[1]*0.99
```

↩ Reply

0




The background features abstract geometric elements: a large orange shape on the left, a blue curved line at the top, a blue curved line on the left, a blue line at the bottom right, and a light gray circle on the bottom right.

02


NETTOYAGE DES DONNÉES

Train



7	variables
122265	observations

Census



26	variables
3142	observations

Test



3	variables
25080	observations

Sample submission



2	variables
25080	observations

Fichiers de travail

Fichier pour la prédiction des mois à venir

Format du fichier des prédictions à remettre pour la compétition

Census

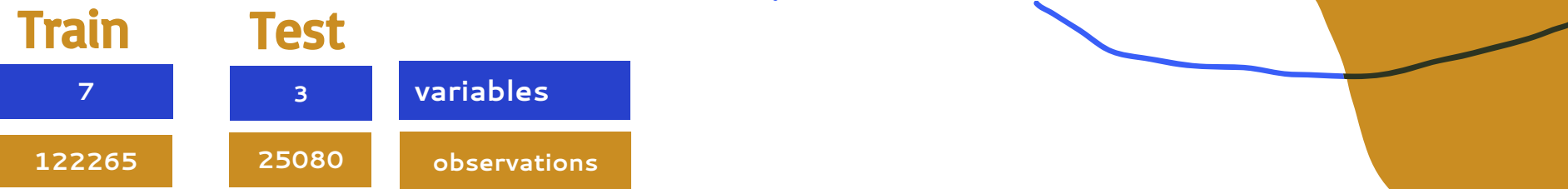


Actions	Nombre	Taille du dataset	Commentaires
Complétion des missing values	12	26x3142	Données complétées avec celle du même comté pour l'année précédente

The background features abstract geometric elements: a large orange shape on the left, a blue curved line at the top, a blue curved line on the left, a blue line at the bottom right, and a light gray circle in the bottom right corner.

03

Manipulations



Actions	Nombre	Taille du dataset raw	Commentaires
Concat Train et Test et Création nouvelle variable, ltest		147345 rows × 8 columns	Train (122265 rows × 7 columns) et Test (25080 rows × 3 columns) ltest : 0/1 Dataframe avec des missings values
Missing values	25080 * 4	147345 rows × 8 columns	County, state, microbusiness, active
Complétion missing values	25080 * 2	147345 rows × 8 columns	County, state
Création de variables quantitatives discrètes (encodage)	5 variables	147345 rows × 13 columns	Year, Month, dcount, county_i, state_i

raw

13	variables
147345	observations

raw	census	
13	26	variables
147345	3142	observations

Actions	Nombre	Taille du dataset	Commentaires
Merge		147345x38	Raw merge census on cfips
Conserver 1 variable de recensement correspondant à l'année de l'observation	20 variables	147345x18	pct_bb, pct_college, pct_foreign, pct_workers, pct_inc (les données de density de l'année A sont calculées avec les données de recensement A-2 ans)
Creation dataframes	2 df	122265 × 18 25080x18	df_train_all df_test_all

df_train_all	df_test_all	
18	18	variables
122265	25080	observations

df_train_all

18

variables

122265

observations

Actions	Nombre	Taille du dataset	Commentaires
Identification des outliers	7 variables	122265 × 18	'Microbusiness_density' : 8746 'Active' : 19183 'Pct_bb' : 2456 'Pct_college' : 2877 'Pct_foreign' : 10097 'Pct_workers' : 3572 'Pct_inc' : 4851
Winsorising	7 variables	122265 × 18	Fonction qui remplace les outliers soit par la valeur upper ou par lower calculés grâce à la méthode IQR; pour chacune des 7 variables

df_train_smoothy

18

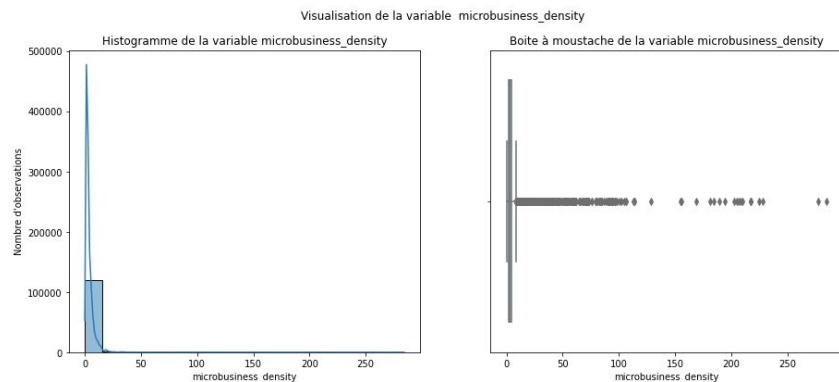
variables

122265

observations

Microbusiness density

Distribution **avant** application Winsorising

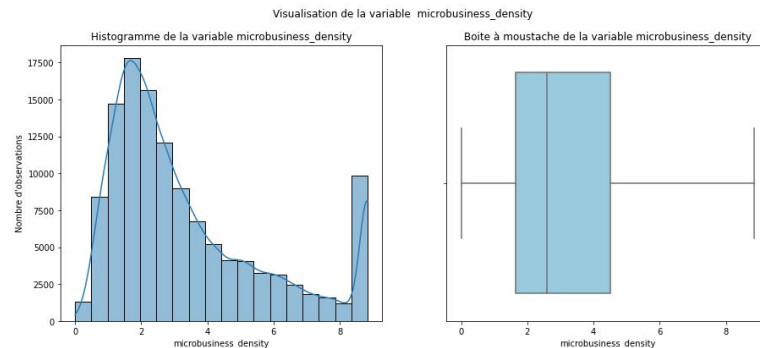


Indicateurs de distribution

count	122265.000000
mean	3.817671
std	4.991087
min	0.000000
25%	1.639344
50%	2.586543
75%	4.519231
max	284.340030

Name: microbusiness_density, dtype: float64

Distribution **après** application Winsorising



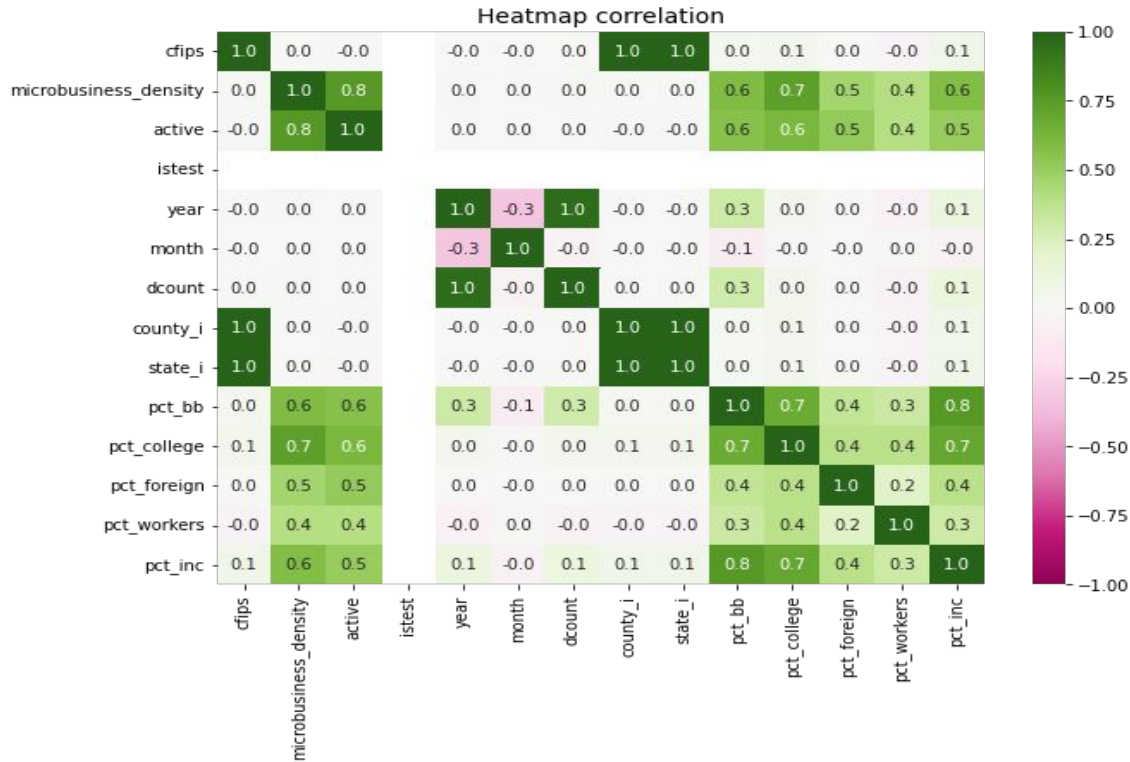
Indicateurs de distribution

count	122265.000000
mean	3.378144
std	2.349475
min	0.000000
25%	1.639344
50%	2.586543
75%	4.519231
max	8.839061

Name: microbusiness_density, dtype: float64

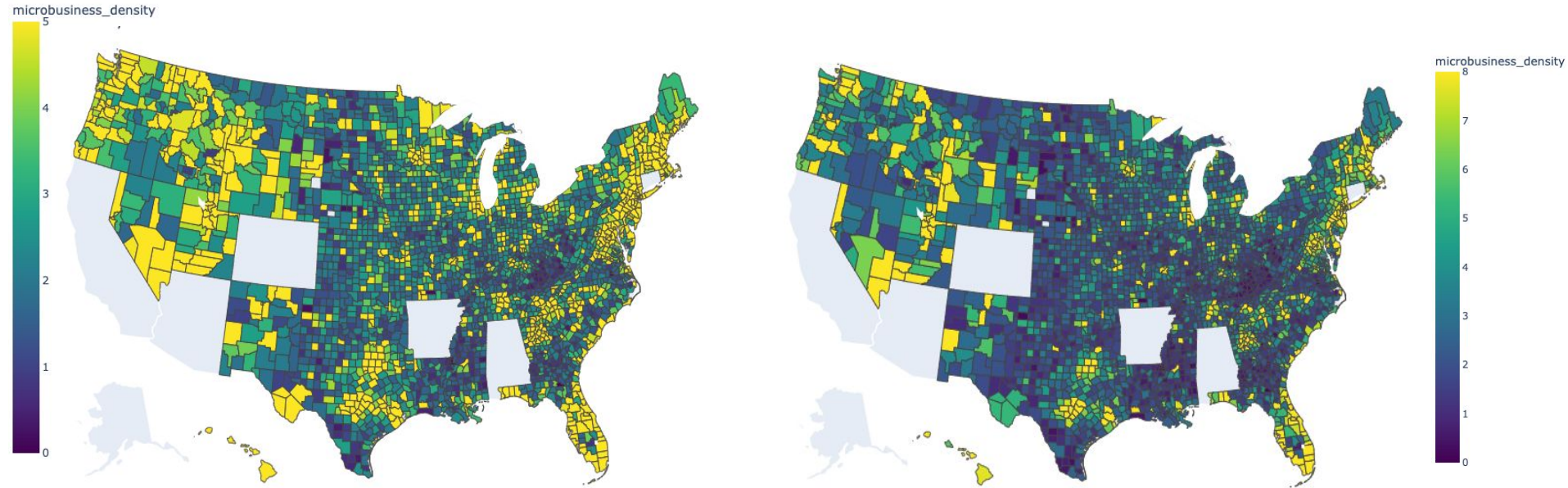
- Asymétrie à droite (positive)
- Données comprises entre 1.6 et 4.5 à 50%

Heatmap correlation



Obtenu avec les données lissées par winsorising

Microbusiness density



Obtenus avec les données lissées par winsorising

The background features abstract geometric elements: a large orange shape on the left, a blue curved line at the top, a blue curved line on the left, a blue line at the bottom right, and a light gray circle on the bottom right.

04

FEATURE ENGINEERING

Création de variables

Df_train_smoothy trié par date décroissante :

density_shift_n avec n allant de 1 à 11	11	Obtenues par un shift de n des valeurs de la variable microbusiness density
dif_n avec n allant de 1 à 11	11	Obtenues par différence de la microbusiness à M avec density_shift_n

Complétion des missing values

Missing values density_shift_n	$3142 * \sum n$	Valeurs complétées par la médiane des density_shift par county
--------------------------------	-----------------	--

df_train_new_feat_smooth

40	variables
122265	observations

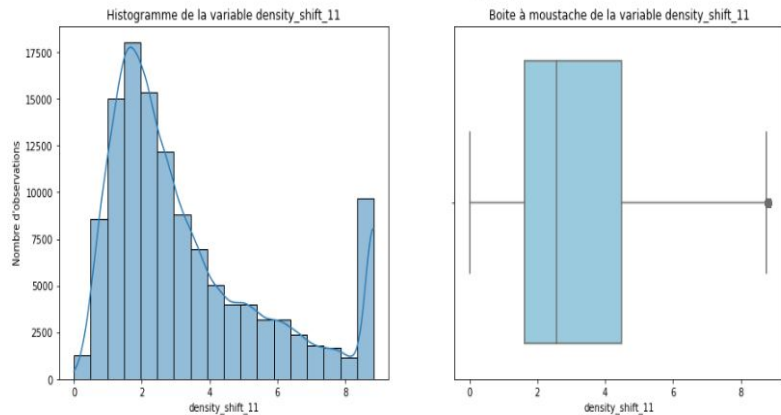
The background features abstract geometric elements: a large orange shape on the left, a blue curved line at the top, a blue curved line on the left, a blue line at the bottom right, and a light gray circle on the bottom right.

05

EXPLORATION DES DONNÉES

Distribution density shift

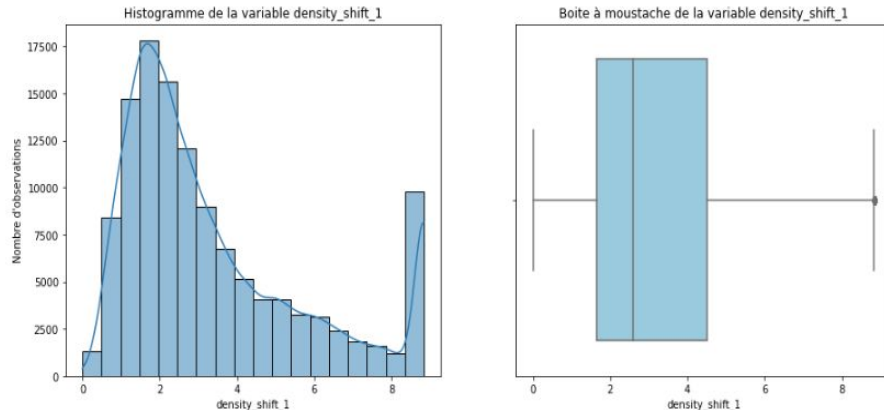
Visualisation de la variable density_shift_11



Indicateurs de distribution

```
count    122265.000000
mean       3.355573
std        2.344294
min         0.000000
25%         1.625677
50%         2.568728
75%         4.470059
max         8.839061
Name: density_shift_11, dtype: float64
```

Visualisation de la variable density_shift_1

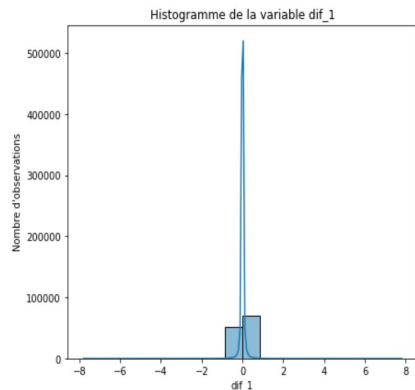


Indicateurs de distribution

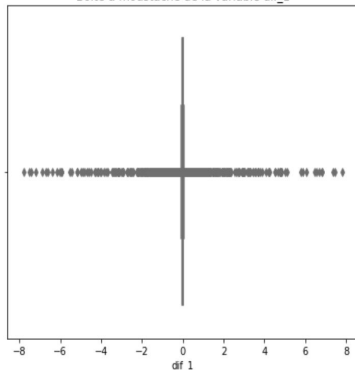
```
count    122265.000000
mean       3.375462
std        2.348568
min         0.000000
25%         1.638081
50%         2.584071
75%         4.516003
max         8.839061
Name: density_shift_1, dtype: float64
```

Distribution dif et log dif

Visualisation de la variable dif_1



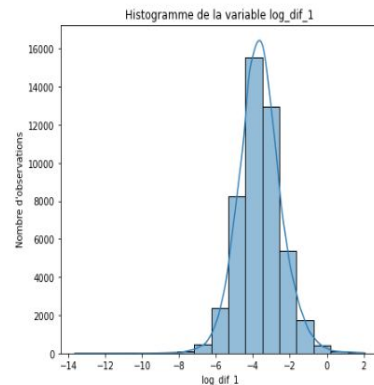
Boite à moustache de la variable dif_1



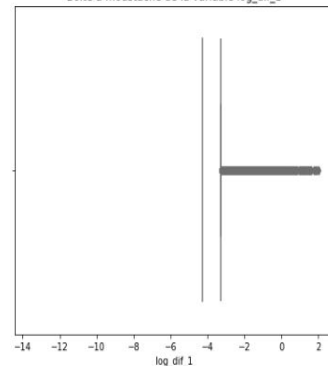
Indicateurs de distribution

```
count    122265.000000
mean      -0.002682
std        0.179233
min       -7.805576
25%       -0.023214
50%        0.000000
75%        0.017531
max        7.797828
Name: dif_1, dtype: float64
```

Visualisation de la variable log_dif_1



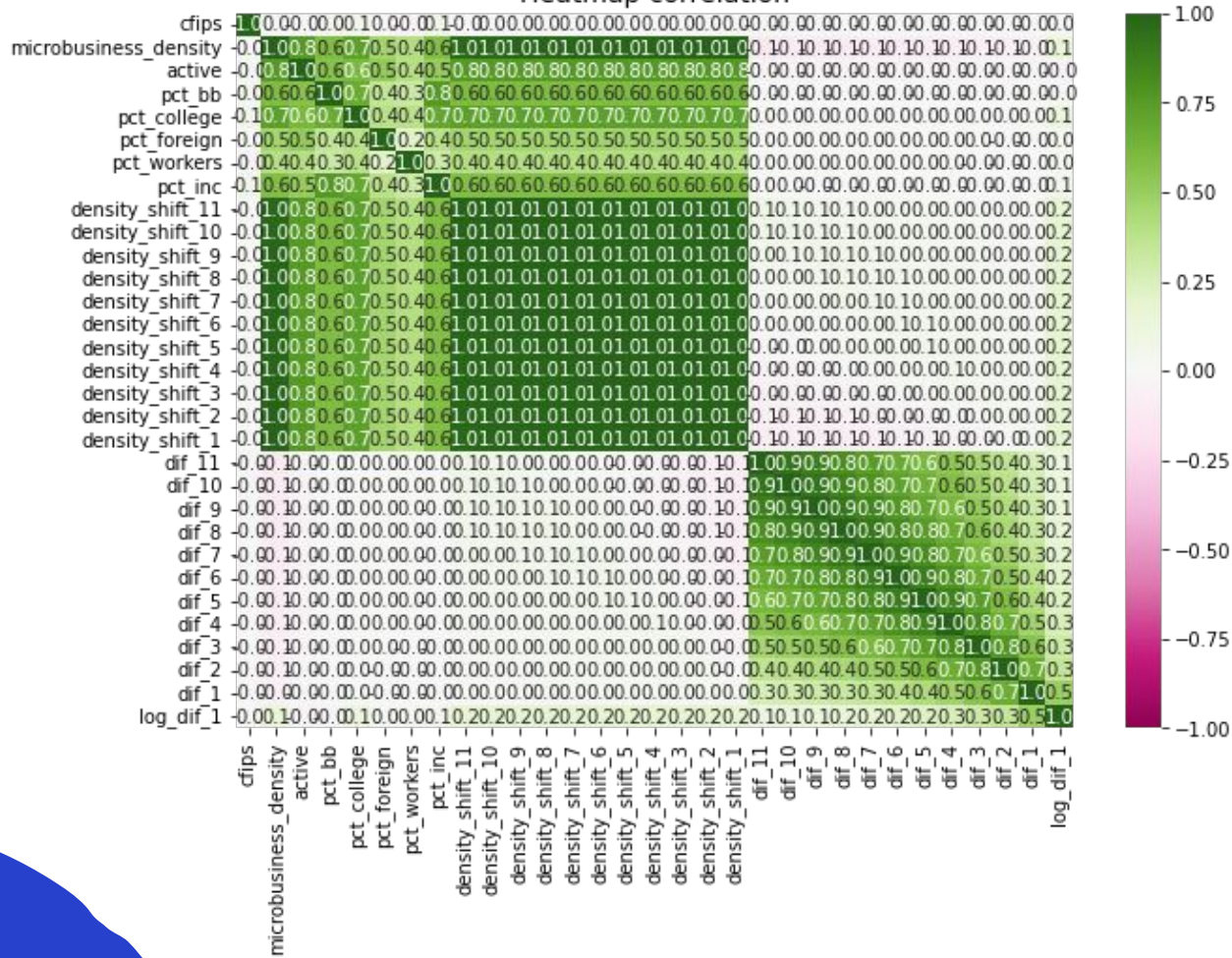
Boite à moustache de la variable log_dif_1



Indicateurs de distribution

```
count    6.835700e+04
mean      -inf
std        NaN
min       -inf
25%        NaN
50%       -4.265204e+00
75%       -3.276296e+00
max        2.053845e+00
Name: log_dif_1, dtype: float64
```


Heatmap correlation



Variables retenues

df_pred_density

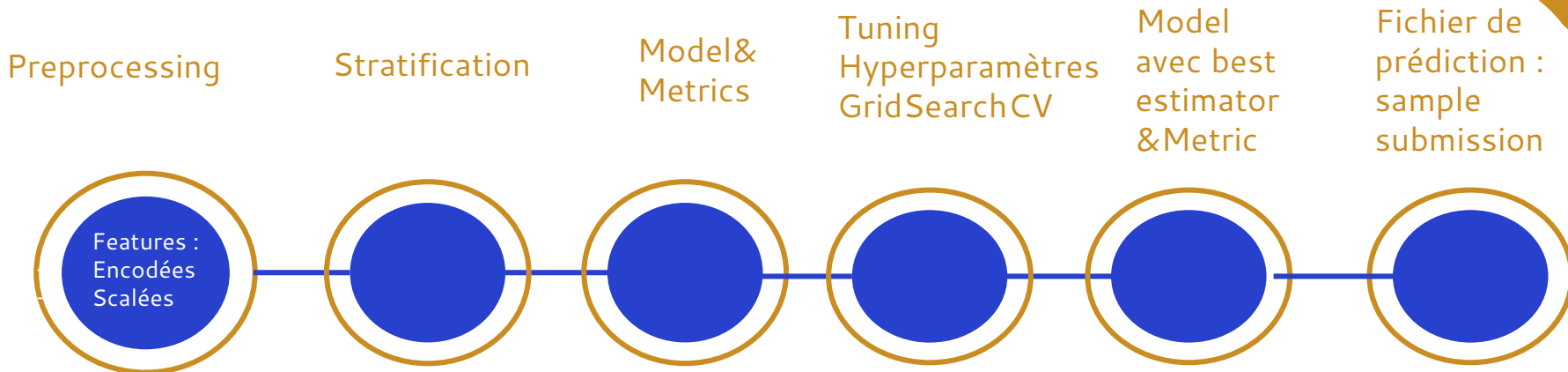
Type de variable	Noms des variables
Période – variables encodées	Year, month
Données géographique du comté	county_i, state_i, dcount
Données de recensement en % (revenu en US dollar ajusté à l'inflation)	pct_bb, pct_college, pct_foreign, pct_workers, pct_inc
Données de densité des microentreprises "shiftées"	Density_shift_n n allant de 1 à 11



06

Méthodologie et Objectifs

Méthodologie et Objectifs



Predictions :

Microbusiness density mensuel
par comté du 01-11-2022 au 01-06-2023

Métrique à minimiser :

SMAPE

The background features abstract geometric elements: a large mustard yellow shape on the left, a blue curved line at the top, two concentric yellow circles in the top right, a light gray circle in the bottom right, and a blue line curving around the bottom right. The word "PREPROCESSING" is centered in a bold, black, sans-serif font.

PREPROCESSING

df_pred_density

18

variables

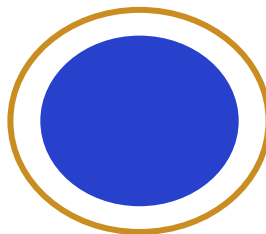
122265

observations

Uniquement sur
les **features** non
encodées ie les
variables de
recensement

Preprocessing testé

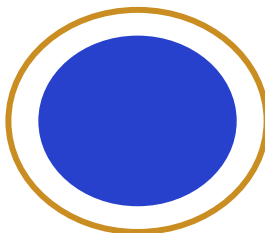
RobustScaler



Après concat avec les autres
features non normalisées :

df_X_density_rs

StandardScaler



df_X_density_Sts

The background features abstract geometric elements: a large mustard yellow shape on the left, a blue curved line at the top, two concentric yellow circles in the top right, a light grey circle in the bottom right, and a blue line curving around the bottom right of the grey circle.

STRATIFICATION

df_pred_density

18

variables

122265

observations

df_X_density_rs

Découpage à la main pour conserver une information continue sur plus d'un an pour tous les comtés (données les plus anciennes)

Split data

Train

85%

`X_density_shift_train`

`y_density_shift_train`

Test

15%

`X_density_shift_test`

`y_density_shift_test`

The background features abstract geometric elements: a large mustard yellow shape on the left, a blue curved line at the top, two concentric orange circles in the top right, a light gray circle in the bottom right, and a blue line curving around the bottom right of the gray circle.

MODELS & METRICS

Modèle Prédiction microbusiness density

	SMAPE	Durée
RANDOMFOREST	1.70	1min9s
GRADIENT BOOSTING	1.72	3min50s
LINEAR REGRESSION	1.58	134 ms

Modèle

Prédiction microbusiness density grid search cv

SMAPE

RANDOMFOREST

1.70

```
'bootstrap': True,  
'max_depth': 80,  
'max_features':  
'auto'
```

GRADIENT BOOSTING

1.72

```
N_estimators : 100  
Learning_rate: 0.1
```

The background features abstract geometric elements: a large mustard yellow shape on the left with a blue curved cutout, a blue wavy line at the top, two concentric yellow circles in the top right, a light grey circle in the bottom right, and a blue line curving around the bottom right of the grey circle.

Sample submission

Sample submission

Prédictions de densité des microentreprises du 01-11-2022 au 01-06-2023 :

La prédiction pour un mois, M :

1. Dataset test avec les infos des comtés pour le mois, M
2. Créer les variables `density_shift_n`, M-1 à M-11, grâce aux densités des microentreprises :
 - du dataset `df_train_new_feat_smooth` (normalisées avec Robuscaler + winsorising)
 - du dataset issu du/des prédict.s précédent.s
3. Concaténer le dataset du point 2 avec le résultat du point 1
4. Faire la prédiction avec le modèle de ML de régression linéaire

Concaténer toutes les prédictions des densités des microentreprises 01-11-2022 au 01-06-2023

Mettre les données au format attendu (`row_id` et `microbusiness_density`)

DÉPÔT DU CODE SUR GITHUB

The screenshot shows the GitHub interface for a repository. At the top, the repository name is 'tatiana-martinez / OPC_8_Competition-Kaggle-Density-Prediction' with a 'Public' badge. Navigation tabs include 'Code', 'Issues', 'Pull requests', 'Actions', 'Projects', 'Wiki', 'Security', 'Insights', and 'Settings'. Below the tabs, there are buttons for 'Go to file', 'Add file', and a green 'Code' button. The repository has 'master' as the selected branch, '1 branch', and '0 tags'. A commit history table follows, showing the latest commit 'tatiana-martinez add' with hash '2bbe0a7' and '7 commits' back. The table lists files: 'data' (first commit, 5 days ago), 'main' (update docstrings, 1 hour ago), '.gitignore' (change gitignore, 5 days ago), 'README.md' (add, 1 hour ago), and 'requirements.txt' (add, 1 hour ago).

tatiana-martinez / OPC_8_Competition-Kaggle-Density-Prediction Public Pin Unwatch

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

master 1 branch 0 tags Go to file Add file <> Code

tatiana-martinez add		2bbe0a7 1 hour ago	🕒 7 commits
📁 data	first commit		5 days ago
📁 main	update docstrings		1 hour ago
📄 .gitignore	change gitignore		5 days ago
📄 README.md	add		1 hour ago
📄 requirements.txt	add		1 hour ago

https://github.com/tatiana-martinez/OPC_8_Competition-Kaggle-Density-Prediction

Perspectives

- Apprentissage (fit) par comté, grâce aux features & microbusiness density
- Intégrer les données de population de plus de 18 ans par comté des USA
- Créer un fichier d'entraînement en supprimant les missing values des variables density shift (au lieu de les remplacer par les médianes)
- Tester l'apprentissage sans les données blacklistées présentes dans le Kernel



MERCI !

Avez-vous des questions?

