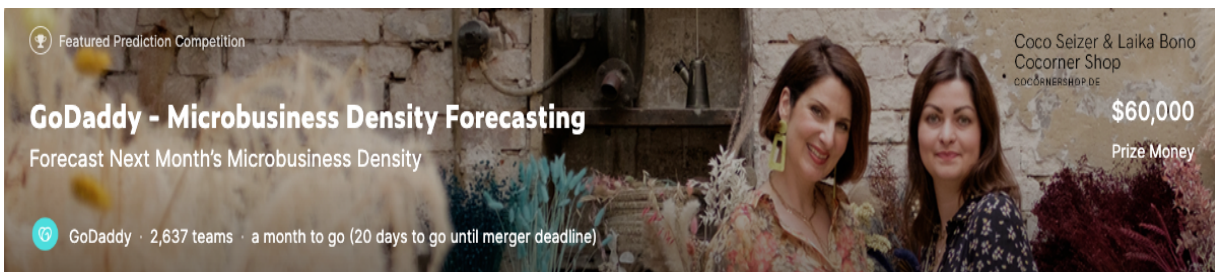


# Openclassrooms projet 8: Compétition Kaggle

## GoDaddy Microbusiness Density Forecasting



**Tatiana Martinez**

Parcours Ingénieur Machine Learning I OpenClassRooms en partenariat avec CentraleSupélec I  
février 2023

<b>INTRODUCTION</b>	<b>2</b>
<b>I. CONTEXTE</b>	<b>2</b>
1. Présentation de la compétition choisie	2
2. Les données	3
3. Remarques	4
<b>II. KERNEL UTILISÉ</b>	<b>5</b>
<b>III. DÉFINITIONS</b>	<b>6</b>
1. Density microbusiness par mois par comté:	6
2. Symmetric Mean Absolute Percentage Error (SMAPE):	6
<b>IV. SIMILITUDES ET DIFFÉRENCES AVEC LE KERNEL</b>	<b>7</b>
1. Similitudes	7
2. Différences	7
<b>V. RÉSULTATS</b>	<b>8</b>
Modèle avec RobustScaler	8
SMAPE	8
Durée du fit	8
<b>CONCLUSION</b>	<b>8</b>

## INTRODUCTION

Le but de ce projet est de participer à une compétition Kaggle réelle en cours. Kaggle est une plateforme qui organise des compétitions en data science et qui récompense les meilleurs analystes internationaux.

L'objectif de ce document est de fournir les résultats et conclusions associées à mon projet pour cette compétition.

## I. CONTEXTE

### 1. Présentation de la compétition choisie

**GoDaddy**, “la plus grande plateforme de services au monde pour les entrepreneurs du monde entier”, a lancé un **challenge Kaggle** afin de mesurer la densité des microentreprises dans les comtés américains.

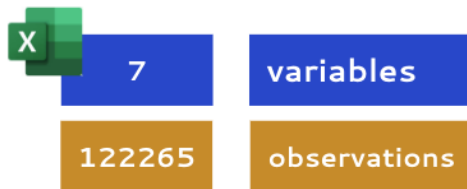
Les micro-entreprises sont souvent trop petites ou trop nouvelles pour apparaître dans les sources de données économiques traditionnelles. Les résultats soumis dans le cadre de cette compétition permettront aux décideurs politiques de mieux connaître la corrélation de l'activité des micro-entreprises avec les autres facteurs extérieurs et ce travail devra leur permettre aussi de créer de bonnes conditions pour celles-ci.

Les participants au challenge ont été invités, 16/12/2022 au 15/03/2023, à proposer des **prédictions de densité de micro-entreprises sur des comtés des USA pour la période de 01-11-2022 au 01-06-2023**. L'**indicateur SMAPE** (Symmetric Mean Absolute Percentage Error) a été retenu pour évaluer les modèles de machine learning qui auront permis de faire les prédictions.

## 2. Les données

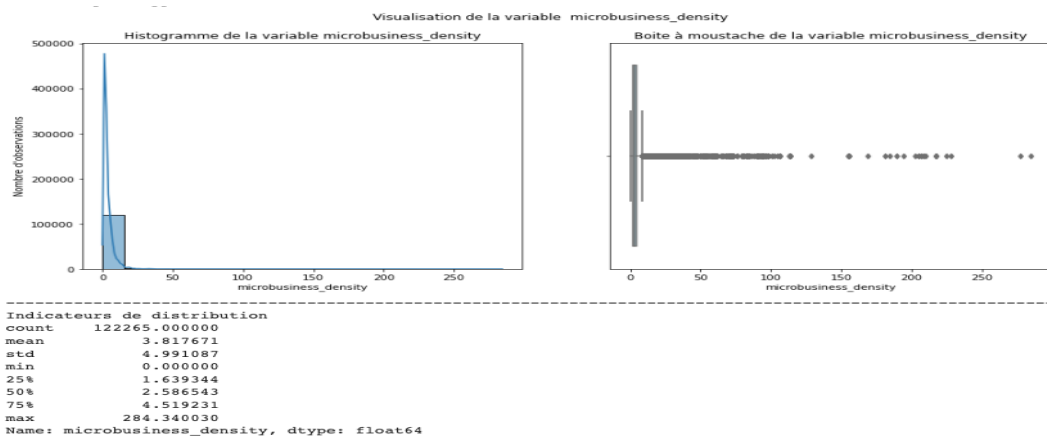
Les organisateurs ont mis à disposition des participants 4 fichiers.

- Fichier train.scv qui fournit la densité mensuelle des microentreprises par comté pour la période du **01-08-2019 au 01-10-2022**



### Train

- row\_id
- cfips
- county\_name
- state\_name
- first\_day\_of\_month
- microbusiness\_density**
- active




- Fichier test.csv qui fournit les références des micro-entreprises sur la période (de prédiction) du **01-11-2022 au 01-06-2023**



### Test

- row\_id
- cfips
- first\_day\_of\_month


- Fichier `census_starter.csv` qui fournit des informations de recensement par comté avec 2 ans de retard par rapport aux mesures de densité de micro-entreprises

	26	variables
	3142	observations

census_starter
<ul style="list-style-type: none"> <li>pct_bb_[year]</li> <li>cfips</li> <li>pct_college_[year]</li> <li>pct_foreign_born_[year]</li> <li>pct_it_workers_[year]</li> <li>median_hh_inc_[year]</li> </ul>

\*year prenant les valeurs de 2017 à 2021

- Fichier `sample_submission` qui fournit des valeurs d'activités des microentreprises. C'est un modèle de format de fichier à soumettre pour concourir à la compétition qui devra fournir les prédictions du **01-11-2022 au 01-06-2023**

	2	variables
	25080	observations

Test
<ul style="list-style-type: none"> <li>row_id</li> <li>microbusiness_density</li> </ul>

### 3. Remarques

Il est possible de s'appuyer sur les kernels partagés par d'autres participants. Il faut prendre soin de citer les kernels utiles.

## II. KERNEL UTILISÉ

Mon projet est basé sur le kernel Better XGB Baseline de Giba publié le 31/12/2022. Il obtient un SMAPE de 1.086.

Son projet suit les étapes suivantes :

1. Préparer le fichier de données de test pour la phase de prédiction
2. Merger les données fournies par le fichier train qui contient notamment la variable de microbusiness\_density avec les données de recensement du fichier census.
3. Identifier les outliers en observant l'évolution du dif de la microbusiness-density. Le dif est obtenu par différence entre microbusiness density à M - microbusiness density à M-1. Il lisse les valeurs en outliers par une moyenne pondérée des valeurs de microbusiness density.
4. Créer une variable target ratio (microbusiness\_density décalé d'un mois) par comté/microbusiness\_density)-1
5. Créer des 11 features, 4 features mbd\_lag\_n obtenues en décalant de n période la variable target (n allant de 1 à 4) et 4 features act\_lag\_n obtenues par différence des données de la variable active et enfin 3 features mbd\_rollmea\_k qui sont les sommes mbd\_lag\_n
6. Créer un modèle XGBRegressor
7. Entraîner le modèle avec les features mbd\_lag\_n, act\_lag\_n, mbd\_rollmea\_k, state\_i, après avoir "blacklisté" des états.

### III. DÉFINITIONS

#### 1. Density microbusiness par mois par comté:

**\*Population de 18 ans et plus du comté**

**Nombre brut de microentreprises  
actives dans le comté sur le mois**

**X 100**

\*Les chiffres de population utilisés pour calculer la densité sont décalés de 2 ans. Ainsi, les chiffres de densité de 2021 sont calculés à partir des chiffres de population de 2019.

#### 2. Symmetric Mean Absolute Percentage Error (SMAPE):

Métrique à réduire; obtenir un résultat le plus proche de 0

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

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

## IV. SIMILITUDES ET DIFFÉRENCES AVEC LE KERNEL

### 1. Similitudes

Les 2 premières étapes du kernel ont été retenues.

### 2. Différences

Les étapes suivantes de mon projet sont différents:

1. Identifier les outliers de 7 variables, par la méthode IQR : 'Microbusiness\_density', 'Active'(le nombre de microentreprises) et les variables de recensement: 'Pct\_bb', 'Pct\_college', 'Pct\_foreign', 'Pct\_workers', 'Pct\_inc'
2. Faire du Winsorising sur les outliers
3. Faire du feature engineering:
  - créer 11 variables de density\_shift\_n issues de la microbusiness\_density décalé de n période en mois et renseigner les valeurs manquantes par la mediane des density\_shift\_n
  - créer 11 variable de dif\_n issues de la différence entre microbusiness\_density et du density\_shift\_n
4. Normaliser les données quantitatives continues avec RobustScaler et StandardScaler
5. Stratification en splittant à la main train et test. Dans le train et le test les données sont ordonnées. Les données les plus anciennes sont dans le train avec 85% des données nettoyées et le reste se trouve dans le test.
6. Créer plusieurs modèles
7. Entraîner le modèle avec les features : active, Year, Month, county\_i, state\_i, dcount, density\_shift\_n et les données de recensement 'Pct\_bb', 'Pct\_college', 'Pct\_foreign', 'Pct\_workers', 'Pct\_inc'
8. Hyperparamétrage tuning avec le Smape à réduire
9. Remplir le fichier de la compétition sample\_submission avec les données de microbusiness\_density prédites par le meilleur modèle



## V. RÉSULTATS

Modèle avec RobustScaler	SMAPE	Durée du fit
Linear Regression	1.58	134 ms
RandomForest	1.70	3min50s
Gradient Boosting	1.72	1min9s

## CONCLUSION

Les résultats sont bons avec un meilleur score obtenu avec le modèle en Linear Regression de 1.58 pour le Smape. Les résultats du kernel sont tout même légèrement meilleurs au niveau du Smape qui est pour rappel de 1.086.

Pour améliorer ce score, voici des axes de travail à tester:

- Apprentissage par comté
- Intégrer les données de population de plus de 18 ans par comté des USA
- Créer un fichier d'entraînement en supprimant les missings values des variables density\_shift (au lieu de les remplacer par les médianes)
- Tester l'apprentissage sans les données blacklistées par Giba

Enfin, le but du projet a été tenu:

- participer à une compétition kaggle en en suivant les règles
- passer par toutes les étapes de l'analyse : récupération, nettoyage des données, analyse exploratoire, création de plusieurs modèles et mesure de leurs performance
- citer les kernels utiles dans les livrables