Modèle de prévision des séismes

Busin Thomas Riboulet-Depret Tristan Van-Duysen Nicolas

Documentation Technique

Introduction

Contexte

Les séismes constituent un risque naturel majeur. L'analyse de données sismiques historiques permet d'identifier des relations entre paramètres géographiques, temporels et physiques des événements.

Objectif du projet

L'objectif est de développer un modèle de régression supervisée capable de :

- prédire la **magnitude** d'un séisme
- prédire la **profondeur** (depth)

à partir des caractéristiques disponibles dans un dataset historique mondial.

Problématique

Comment exploiter efficacement des données hétérogènes (spatiales, temporelles, physiques) pour prédire des variables continues liées aux séismes ?

Données

Description du dataset

- **Source** : Kaggle – *200 Years of Global Major Earthquakes*
- **Période** : 1826–2026
- **Taille** : 17 features & 106077 observations
- **Variables principales** :
    - latitude
    - longitude
    - depth (en km)
    - mag (magnitude)
    - magType (unité utilisé pour la magnitude)
    - time (date et heure de l'événement)
    - updated (dernière fois que les données ont été mise à jour)
    - place (pays)
    - type (type d'événement)
    - nst (nombre de stations sismiques utilisées)
    - gap (le plus grand angle (en degrés) entre deux stations sismiques consécutives)
    - dmin (distance minimum jusqu'à une station sismique)
    - rms (racine carré de la moyenne au carré des résidus du temps de transport)
    - horizontal_error (incertitude de la localisation en km)

- depthError (incertitude de la profondeur en km)
- magError (incertitude de la magnitude)
- magNst (nombre de station contribuant à la détection de la magnitude)

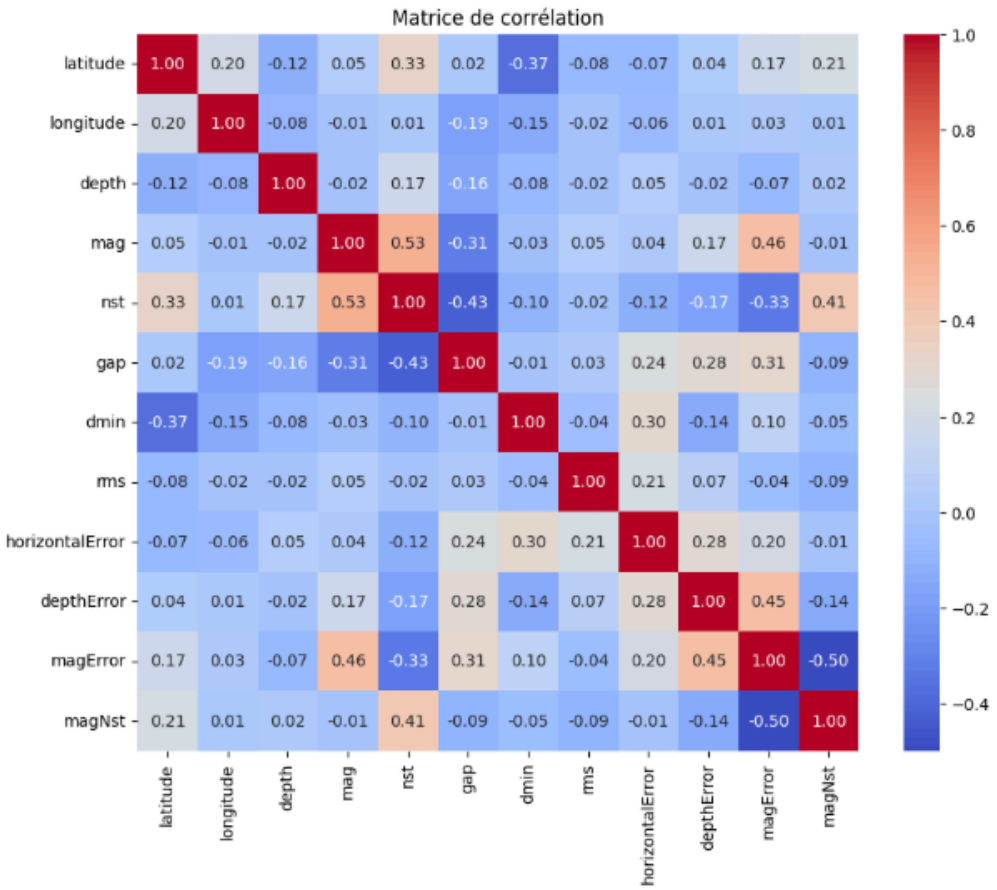
[Lien vers la Datacard]

Exploration des données (EDA)

Graphes

(cf Graphs des données dans l'annexe)

Analyse de corrélation



Observations clés

- Corrélation faible/modérée entre profondeur et magnitude
- Forte dépendance spatiale (latitude/longitude)

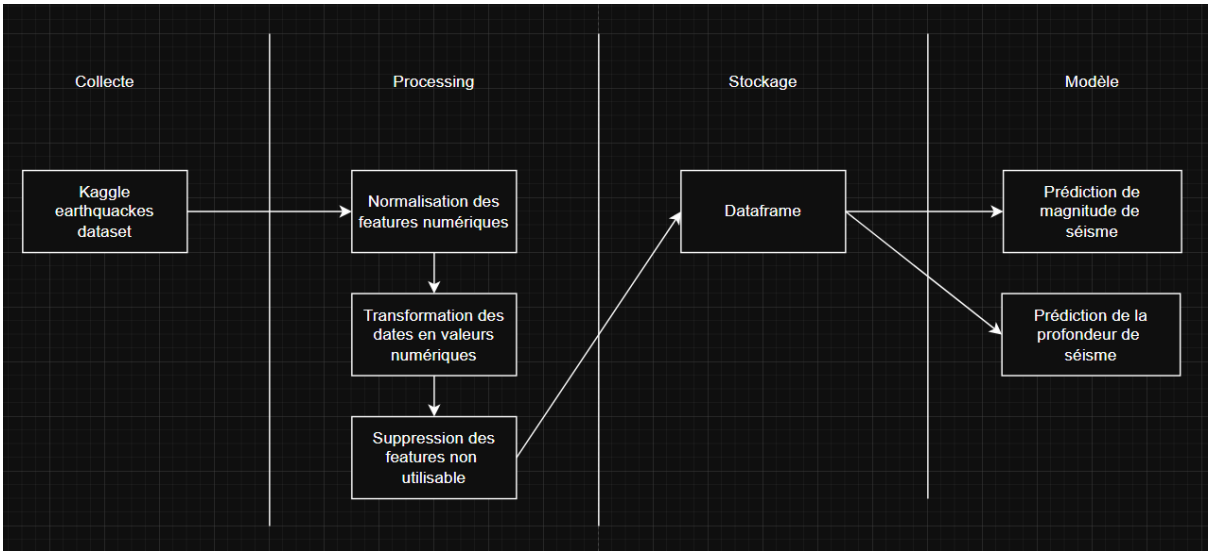
Prétraitement et transformations

- Suppression des valeurs manquantes / aberrantes
- Conversion des dates en variables numériques
- Sélection des features pertinentes
- Normalisation

[Lien vers le notebook]

Architecture du projet

Diagramme de flux des données



Modèle

Choix du modèle

Le modèle XGBoost Regressor a été retenu pour sa robustesse face aux relations non linéaires, sa capacité à capturer efficacement les interactions entre variables et ses performances reconnues sur des données tabulaires.

Variables d'entrée et de sortie

- **Features (X)** :
  - latitude
  - longitude
  - time
  - nst
  - gap

    - dmin
    - rms
    - horizontal_error
    - depthError
    - magError
    - magNst

- **Targets (y)** :
    - mag
    - depth

Entraînement

Les données ont été divisées en un ensemble d'entraînement (80 %) et un ensemble de test (20 %) à l'aide de la méthode "train_test_split()" de la librairie "sklearn". Les performances du modèle sont évaluées à l'aide du coefficient de détermination R^2, qui est la méthode de base de la fonction "score()" de XGBoost. Les tests d'hyperparamètres ainsi que la comparaison avec d'autres modèles seront réalisés après les prochaines itérations du projet.

Résultats

Sur les modèles de base on remarque un score R^2 de 0.62 sur le modèle visant à prédire la magnitude des séismes, et un score R^2 de 0.839 pour le modèle sur la profondeur des séismes.

Évaluation et limites

Performances

Les performances brutes du modules sont mitigées. Le modèle prédisant la magnitude des séismes sous performe et n'est pas fiable en l'état. Une manipulation de données plus précises serait nécessaire. En revanche, le modèle visant à prédire la profondeur semble correct avec un score supérieur à 0.80 sur la première itération.

Limites

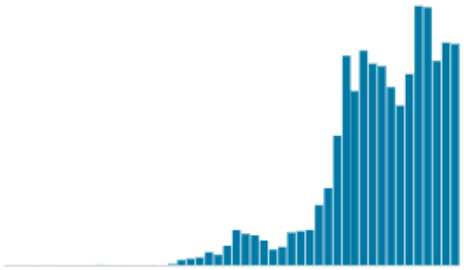
Les données sismiques du dataset sont incomplètes et hétérogènes, en particulier avant le XX^{e} siècle, ce qui peut affecter la qualité des prédictions. Ce projet ne permet pas de réaliser une prédiction déterministe des séismes à court terme, mais uniquement des estimations statistiques basées sur les données trouvées.

Annexes

Graphs des données

time

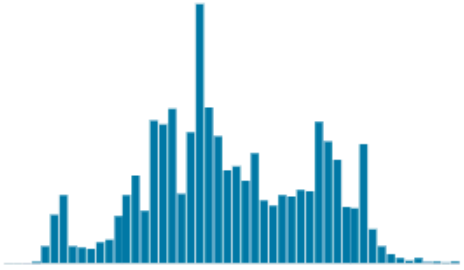
UTC timestamp when the earthquake occurred



Valid	106k	100%
Mismatched	0	0%
Missing	0	0%
Minimum	24Sep27	
Mean	9Nov91	
Maximum	7Jan26	

latitude

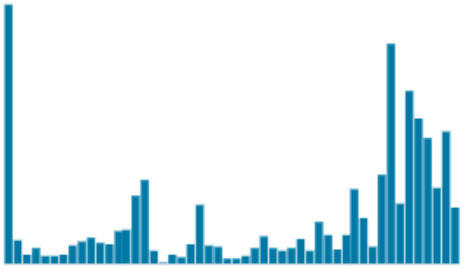
Latitude of the earthquake epicenter



Valid	106k	100%
Mismatched	0	0%
Missing	0	0%
Mean	3.79	
Std. Deviation	30.3	
Quantiles	-77.1	Min
	-17.7	25%
	-0.57	50%
	30.2	75%
	87.4	Max

longitude

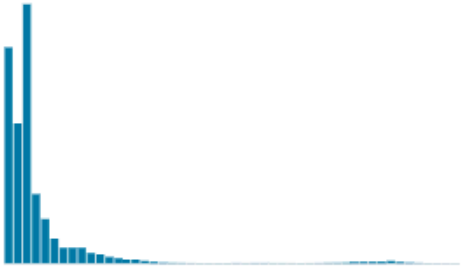
Longitude of the earthquake epicenter



Valid	106k	100%
Mismatched	0	0%
Missing	0	0%
Mean	40.3	
Std. Deviation	122	
Quantiles	-180	Min
	-72.4	25%
	99.1	50%
	143	75%
	180	Max

depth

Depth of the earthquake in kilometers



Valid	105k	99%
Mismatched	0	0%
Missing	593	1%
Mean	61.6	
Std. Deviation	108	
Quantiles	-4	Min
	12	25%
	33	50%
	50.3	75%
	700	Max

place

Human-readable location description

South Sandwich Islands region	2%
Kermadec Islands region	1%
Other (102111)	96%

Valid	106k	100%
Mismatched	0	0%
Missing	222	0%
Unique	64.3k	
Most Common	South Sand...	2%

type

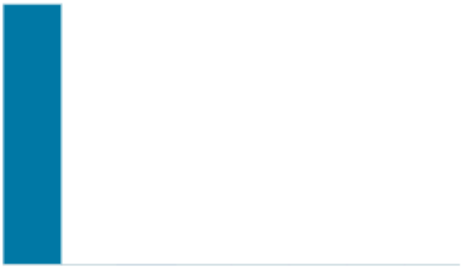
Event type (earthquake)

earthquake	100%
nuclear explosion	0%
Other (68)	0%

Valid	106k	100%
Mismatched	0	0%
Missing	0	0%
Unique	7	
Most Common	earthquake	100%

horizontalError

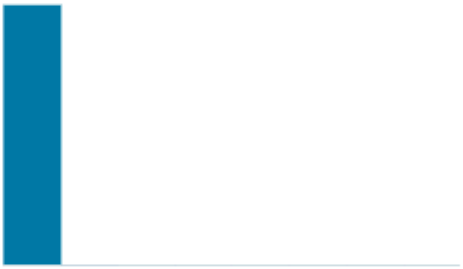
Horizontal location uncertainty (km)



Valid	20.0k	19%
Mismatched	0	0%
Missing	86.1k	81%
Mean	7.83	
Std. Deviation	8.64	
Quantiles	0	Min
	6.03	25%
	7.5	50%
	9.2	75%
	778	Max

depthError

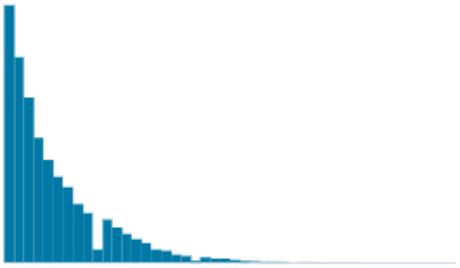
Depth uncertainty (km)



Valid	54.4k	51%
Mismatched	0	0%
Missing	51.6k	49%
Mean	7.98	
Std. Deviation	10.7	
Quantiles	-1	Min
	1.98	25%
	4.9	50%
	9.6	75%
	1.09k	Max

mag

Earthquake magnitude



Valid	106k	100%
Mismatched	0	0%
Missing	0	0%
Mean	5.45	
Std. Deviation	0.49	
Quantiles	5	Min
	5.1	25%
	5.3	50%
	5.7	75%
	9.5	Max

A magType

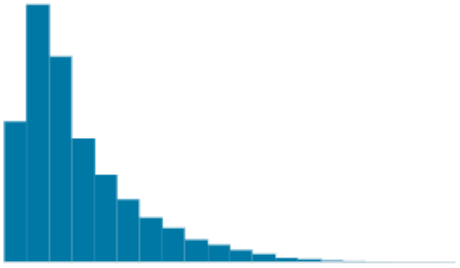
Magnitude scale used (Mw, Mb, ML, etc.)

mb	39%
mw	25%
Other (37686)	36%

Valid	106k	100%
Mismatched	0	0%
Missing	0	0%
Unique	28	
Most Common	mb	39%

nst

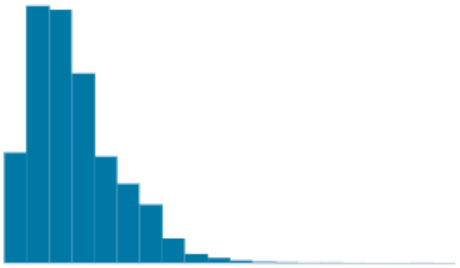
Number of seismic stations used



Valid	31.3k	30%
Mismatched	0	0%
Missing	74.7k	70%
Mean	158	
Std. Deviation	126	
Quantiles	0	Min
	68	25%
	118	50%
	210	75%
	934	Max

gap

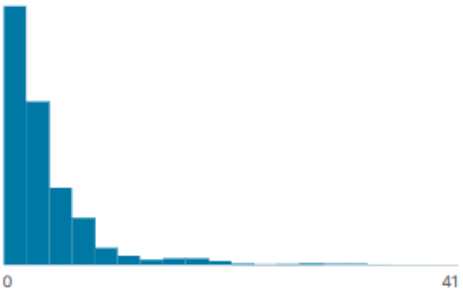
Azimuthal gap indicating station coverage (degrees)



Valid	42.0k	40%
Mismatched	0	0%
Missing	64.1k	60%
Mean	63.1	
Std. Deviation	38.6	
Quantiles	6.5	Min
	36	25%
	54.7	50%
	80.7	75%
	360	Max

dmin

Minimum distance to nearest seismic station (degrees)



Valid	21.4k	20%
Mismatched	0	0%
Missing	84.7k	80%
Mean	4.24	
Std. Deviation	5.21	
Quantiles	0	Min
	1.27	25%
	2.53	50%
	5.01	75%
	41	Max

rms

Root mean square of travel-time residuals

Valid	74.6k	70%
Mismatched	0	0%
Missing	31.5k	30%
Mean	0.96	
Std. Deviation	0.37	
Quantiles	-1	Min
	0.81	25%
	0.97	50%
	1.1	75%
	69.3	Max

A id

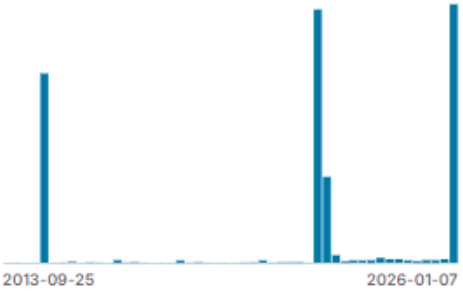
Unique earthquake event identifier

106077
unique values

Valid	106k	100%
Mismatched	0	0%
Missing	0	0%
Unique	106k	
Most Common	us7000rnbi	0%

updated

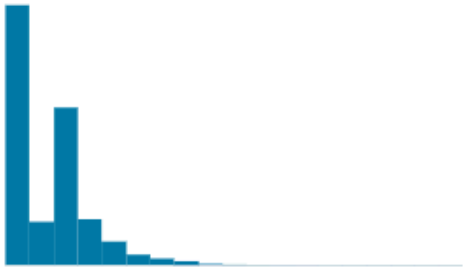
Last update timestamp of the event



Valid	106k	100%
Mismatched	0	0%
Missing	0	0%
Minimum	25Sep13	
Mean	21Sep21	
Maximum	7Jan26	

magError

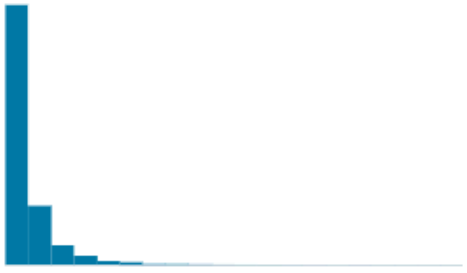
Magnitude uncertainty



Valid	36.6k	34%
Mismatched	0	0%
Missing	69.5k	66%
Mean	0.17	
Std. Deviation	0.15	
Quantiles	0	Min
	0.06	25%
	0.1	50%
	0.21	75%
	1.84	Max

magNst

Number of stations contributing to magnitude



Valid	42.4k	40%
Mismatched	0	0%
Missing	63.7k	60%
Mean	54.9	
Std. Deviation	83	
Quantiles	0	Min
	12	25%
	27	50%
	61	75%
	1.03k	Max

A status

Review status (automatic or reviewed)

reviewed	100%
automatic	0%

Valid	106k	100%
Mismatched	0	0%
Missing	0	0%
Unique	2	
Most Common	reviewed	100%