

Cursus Ingénieur Machine Learning

-

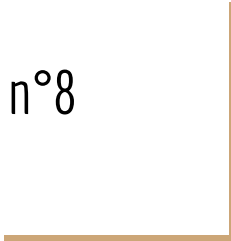
Vincent Jugé

-

Soutenance Projet n°8

-

11/2022





Participez à une compétition Kaggle




Choix du domaine - TSF

- Le Time Series Forecasting (TSF) est un champ du machine learning qui consiste à prédire des valeurs futures d'un phénomène temporel.
- Ce domaine couvre des applications très variées : météo, logistique, traitement du signal, sismologie, économie, finance, ...


Challenge

- Store Sales - Time Series Forecasting
 - <https://www.kaggle.com/competitions/store-sales-time-series-forecasting/overview>
 - Pas d'argent à gagner - compétition pour les débutants sur Kaggle
- *Goal of the Competition*
 - *In this “getting started” competition, you’ll use time-series forecasting to forecast store sales on data from Corporación Favorita, a large Ecuadorian-based grocery retailer.*
 - *Specifically, you'll build a model that more accurately predicts the unit sales for thousands of items sold at different Favorita stores. You'll practice your machine learning skills with an approachable training dataset of dates, store, and item information, promotions, and unit sales.*

 GettingStarted Prediction Competition

Store Sales - Time Series Forecasting

Use machine learning to predict grocery sales

 Kaggle · 715 teams · Ongoing

Overview

Data

Code

Discussion

Leaderboard

Rules

Team

Submissions

Submit Predictions

...

Overview

Description


Evaluation

Frequently Asked Questions

Goal of the Competition

In this “getting started” competition, you’ll use time-series forecasting to forecast store sales on data from Corporación Favorita, a large Ecuadorian-based grocery retailer.

Specifically, you’ll build a model that more accurately predicts the unit sales for thousands of items sold at different Favorita stores. You’ll practice your machine learning skills with an approachable training dataset of dates, store, and item information, promotions, and unit sales.

 **Get Started**

We highly recommend the [Time Series course](#), which walks you through how to make your first submission. The lessons in this course are inspired by winning solutions from past Kaggle time series forecasting competitions.

Context

Forecasts aren’t just for meteorologists. Governments forecast economic growth. Scientists attempt to predict the future population. And businesses forecast product demand—a common task of professional data scientists. Forecasts are especially relevant to brick-and-mortar grocery stores, which must dance delicately with how much inventory to buy. Predict a little over, and grocers are stuck with overstocked, perishable goods. Guess a little under, and popular items quickly sell out, leading to lost revenue and upset customers. More accurate forecasting, thanks to machine learning, could help ensure retailers please customers by having just enough of the right products at the right time.

Current subjective forecasting methods for retail have little data to back them up and are unlikely to be automated. The problem becomes even more complex as retailers add new locations with unique needs, new products, ever-transitioning seasonal tastes, and unpredictable product marketing.

Potential Impact

If successful, you’ll have flexed some new skills in a real world example. For grocery stores, more accurate forecasting can decrease food waste related to overstocking and improve customer satisfaction. The results of this ongoing competition, over time, might even ensure your local store has exactly what you need the next time you shop.

Implementation

Dataset

- Le dataset est composés de plusieurs sources de données
 - Données numériques et catégorielles
 - > 3'000'000 lignes
 - Données horodatées 01/01/2013 au 31/08/2017
 - On cherche à prédire le volume de ventes entre 16/08/2017 et 31/08/2017
- Des données complémentaires sont présentes:
 - Promotion, événements exceptionnels, cours du pétrole
- Le résultat à produire ou sous la forme d'un fichier "clé - valeur"

date	id	store_nbr	family	sales	onpromotion	family_cat
2013-01-01	0	1	AUTOMOTIVE	0.0	0	0
2013-01-01	1194	42	CELEBRATION	0.0	0	6
2013-01-01	1193	42	BREAD/BAKERY	0.0	0	5
2013-01-01	1192	42	BOOKS	0.0	0	4
2013-01-01	1191	42	BEVERAGES	0.0	0	3
...
2017-08-15	2999695	25	POULTRY	172.517	0	28
2017-08-15	2999694	25	PLAYERS AND ELECTRONICS	3.0	0	27
2017-08-15	2999693	25	PET SUPPLIES	3.0	0	26
2017-08-15	2999704	26	BOOKS	0.0	0	4
2017-08-15	2999107	1	BABY CARE	0.0	0	1

3008016 rows x 6 columns

Dataset

- Variables retenues
 - Date, ID Store, Family, Sales
- Données Multi Variées
 - Complexité !
 - Besoin de splitter par ID Store et Family
 - Besoin d'avoir un modèle compatible Multi Variate
- Données manquantes
 - 25/12 manque
 - Traitement nécessaire : les modèles utilisés necessitent une continuité des series
- Prévisions
 - 16 jours, 54 stores, 33 familles = 28512 points à prédire

Choix du Modèle

- GluonTs
 - Librairie avec approche probabiliste
 - Modèle Deep Learning
 - Donne de bons résultats - out of the box
- Autres approches envisagés
 - Prophet (facebook.github.io/prophet)
 - XGBoost / LightGBM

Focus GluonTS : ts.gluon.ai

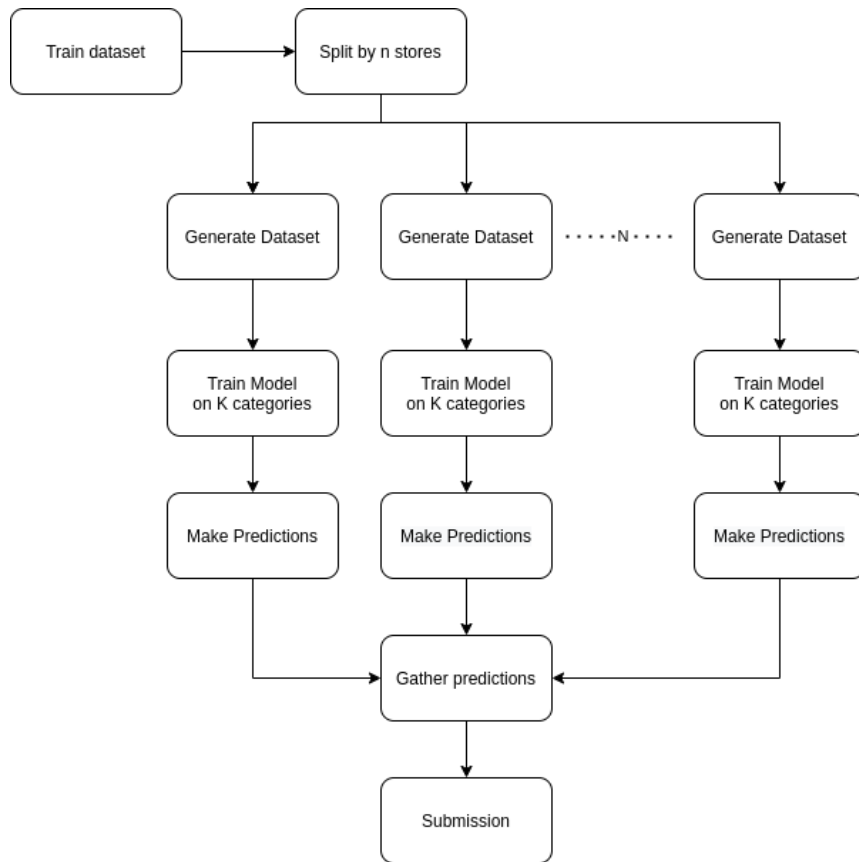
- Modèles à base de Deep Learning
 - RNNs, LSTM, Transformers, ...
 - Univariés, Multivariés
 - + 20 modèles pré-entraînés
- Open Source
- Format de données spécifique
 - Nécessite un formattage
 - Contraintes de continuité des séries temporelles

GluonTs – Format de données

- Splitte le dataset en 54 dataset correspondants à chaque stores
- PandasDataset permet de convertir depuis un DataFrame
 - <https://ts.gluon.ai/stable/api/gluonts/gluonts.dataset.pandas.html>
 - Valeurs:
 - Target : feature qu'on souhaite prédire (sales)
 - Freq: fréquence de la time series (ici 1 jour)
 - Item_id : feature catégorielle pour différencier les time series (family)

Workflow

- Détail du workflow implémenté pour traiter un problème multivarié

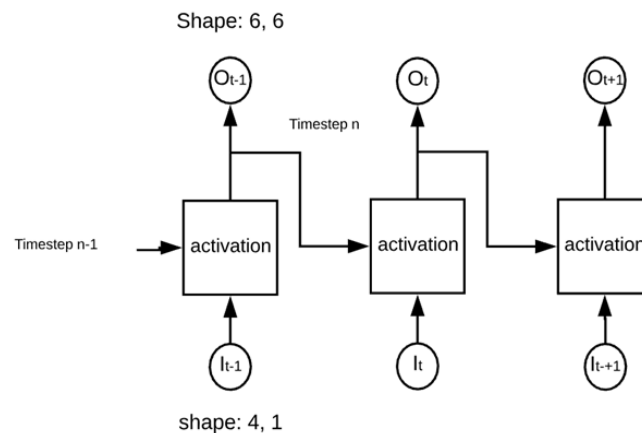


Modèle retenu / Métriques

- Utilisation de DeepAR (RNN)
 - <https://www.sciencedirect.com/science/article/pii/S0169207019301888?via%3Dihub>
 - Auto regressive RNN
- Hyper Parametres
 - 5 epochs
 - Learning rate: 10^{-3}
 - Batches par epoch: 10
 - Early stopping: 2
- Metric retenu: RMSE
 - Minimum ~285 en moyenne sur l'ensemble des prévisions, en fonction des couples ID Store / Famille de produit

Focus DeepAR / RNN

- Les RNN sont dérivés des Feedforward Neural Network (single-layer / multi layers perceptron)
- Les RNN retiennent une mémoire des données qu'ils ont déjà traitées, et donc peuvent apprendre des itérations précédentes.
- Développés initialement pour le NLP



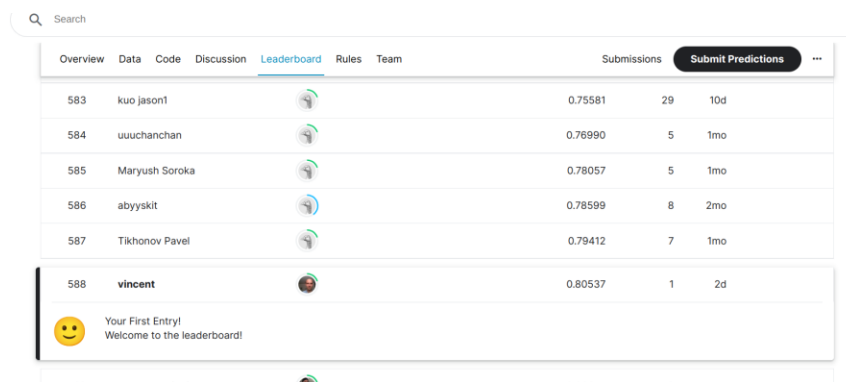
Résultats et Suite

- Soumission des résultats

- Score 0.8
- 588eme sur +700
- <https://www.kaggle.com/code/vincentjuge/notebookcfaf1d1866>

- Marges de manoeuvre

- Data : Ajouter les evenements exceptionnels
- Data : scaler / nettoyer les données (IQR) ?
- Modèle : ajuster les hyper parametres / grid search
- Modèle : remplacer DeepAR par un modèle LSTM ? (LSTnet)
- Modèle : changer complètement l'approche pour du Gradient Boosting?

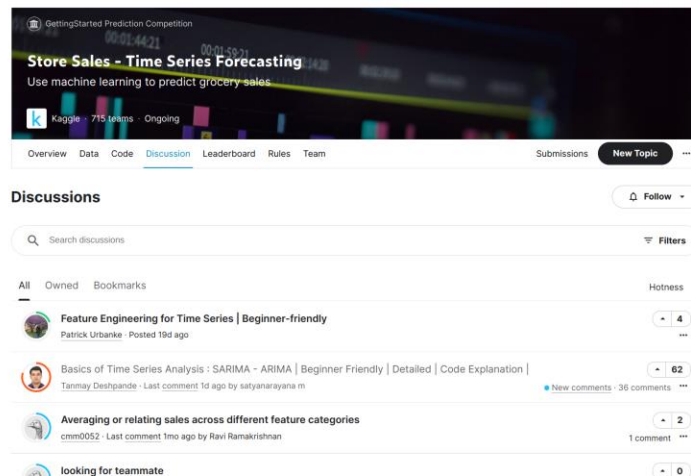


The screenshot shows a Kaggle leaderboard interface. At the top, there are tabs for Overview, Data, Code, Discussion, Leaderboard (selected), Rules, and Team. On the right, there are links for Submissions and a Submit Predictions button. The table lists several users with their scores, ranks, and time since last update. The user 'vincent' is highlighted at the bottom of the table with a score of 0.80537, rank 1, and 2 days since last update. Below the table, a message says 'Your First Entry! Welcome to the leaderboard!' with a smiley face icon.

Rank	User	Score	Rank	Time
583	kuo jason1	0.75581	29	10d
584	uuuchanchan	0.76990	5	1mo
585	Maryush Soroka	0.78057	5	1mo
586	abyyskit	0.78599	8	2mo
587	Tikhonov Pavel	0.79412	7	1mo
588	vincent	0.80537	1	2d

Conclusion

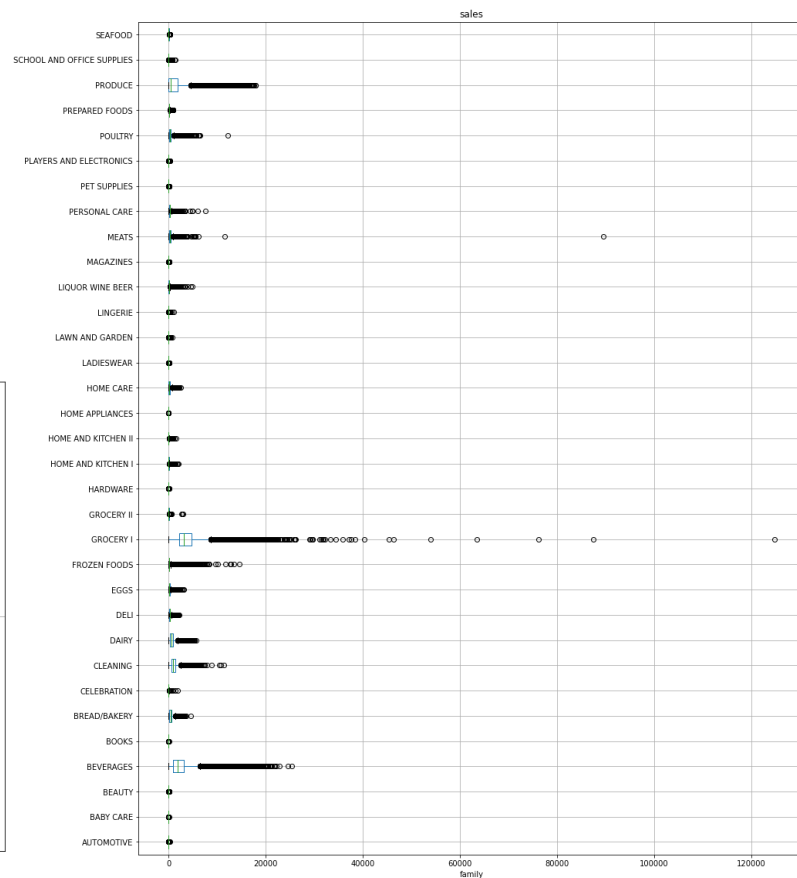
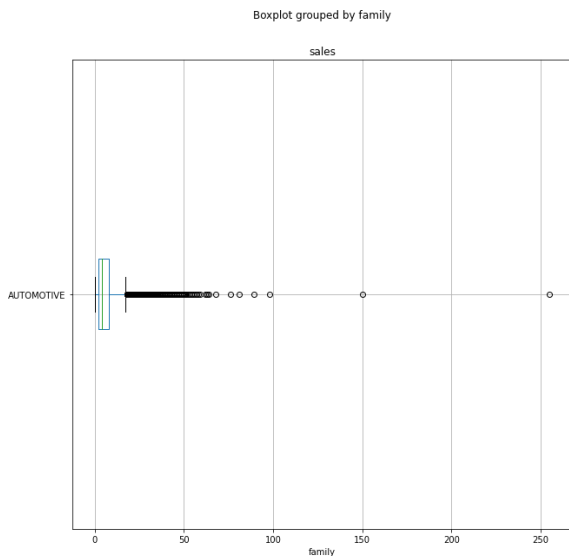
- Première compétition !
- Résultats encourageants
 - Participer à d'autres compétitions futures
- Grande Communauté
 - Permet de découvrir de nouvelles approches
 - Par ex. XGBoost
- Regrets
 - Développement en local – kernel kaggle est trop lent / cher
 - Manque de temps pour optimiser
 - On ne peut pas voir les meilleurs notebook pour comprendre et apprendre



Compléments

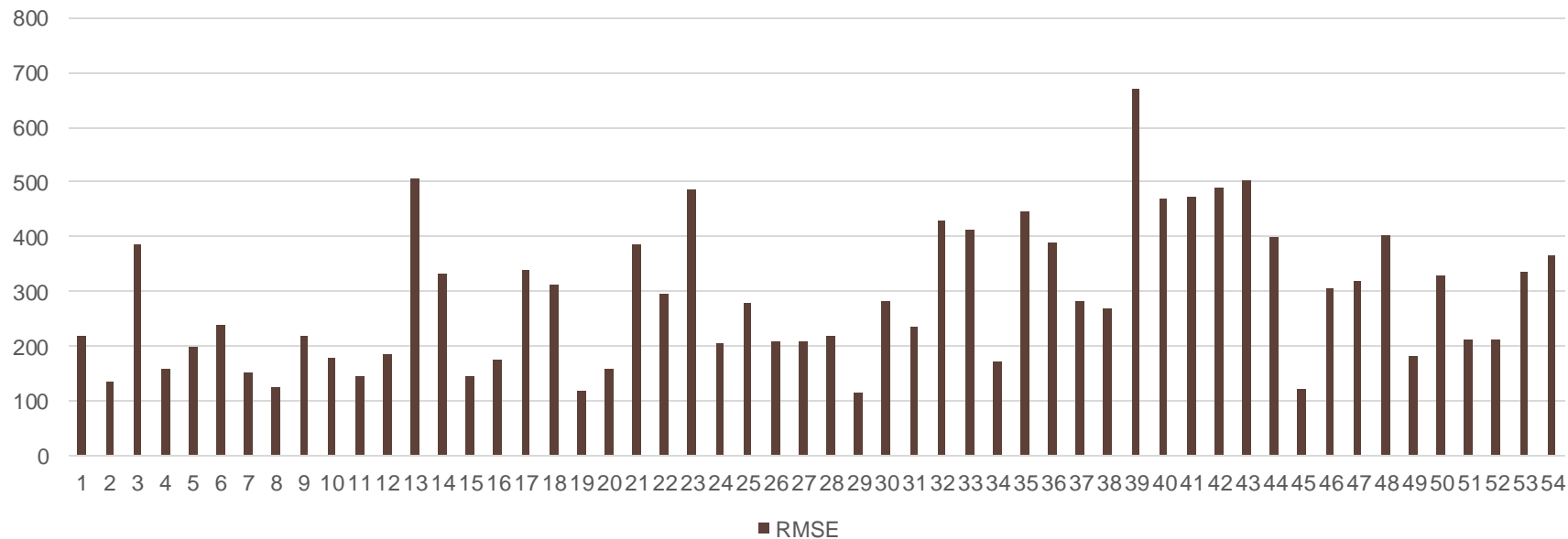
Scaling des données

- Outliers présents



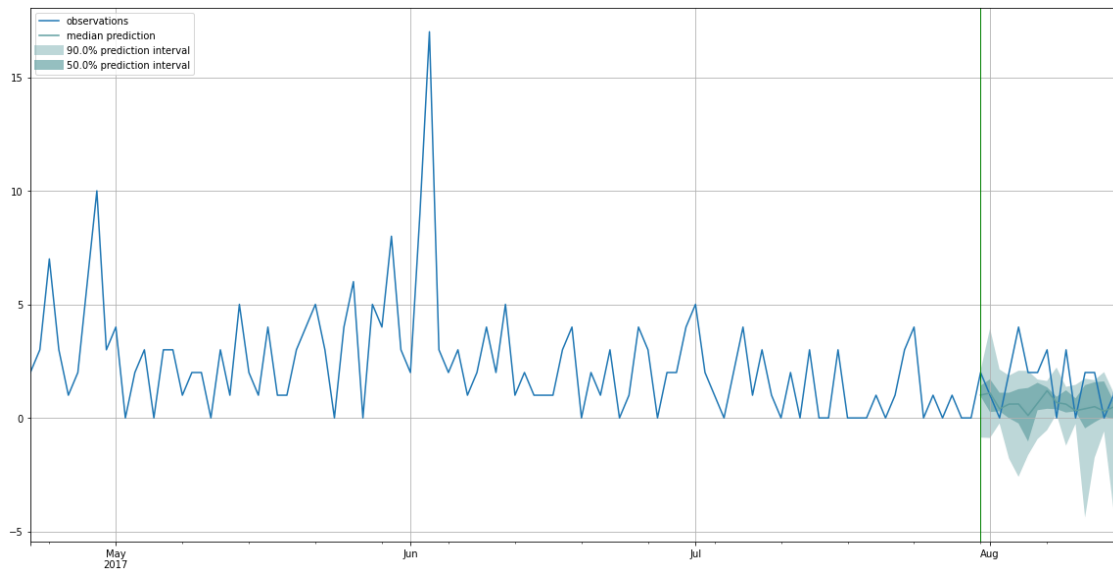
RMSE

RMSE for Each Store



Predictions

- Predictions pour un store
- RMSE = 136












Liens utiles

- Website avec les notebooks : <https://vjuge.github.io/oc-impl>
- Repo Git : <https://github.com/vjuge/oc-impl>
 - Notebook : <https://github.com/vjuge/oc-impl/blob/master/modules/P8/module-p8.ipynb>
- Notebook
Kaggle: <https://www.kaggle.com/code/vincentjuge/notebookcfae1d1866/edit/run/112055289>

Last Minute Update

- Including Oil data, gives a better score of 0.75 (-5% gain)

Overview	Data	Code	Discussion	Leaderboard	Rules	Team	Submissions		Submit Predictions	...
592	Alexandre Fleutelot			0.70682	1	2mo				
593	Pasha Biglarzadeh			0.71926	1	16d				
594	Alex Zelentsov			0.72213	2	1mo				
595	Chutian Sun			0.72402	1	16d				
596	kuoliu0316			0.72402	1	16d				
597	Richyyy			0.72402	1	7d				
598	vincent			0.75385	9	1s				
<div> Your Best Entry! Your most recent submission scored 0.75385, which is an improvement of your previous score of 0.77840. Great job!</div>							<div>Tweet this</div>			
599	...			0.75402	1	...				