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 dans valeurs futures d'un phénomène observé.

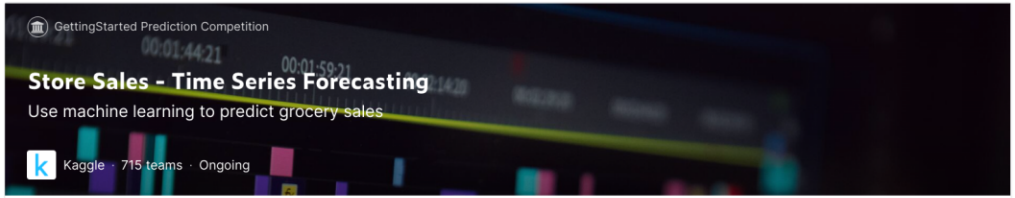
- Ce domaine couvre des applications très variées : météo, logistique, traitement du signal, sismologie, économie, finance, ...

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.*



Overview

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 plusieures 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 en fonction de:
    - ID du magasin
    - Catégorie de produits
  - Des données complémentaires présentes:
    - Promotion, évênements exceptionnels, cours du pétrole
- Le résultat à produire ou sous la forme d'un fichier "clé - valeur"

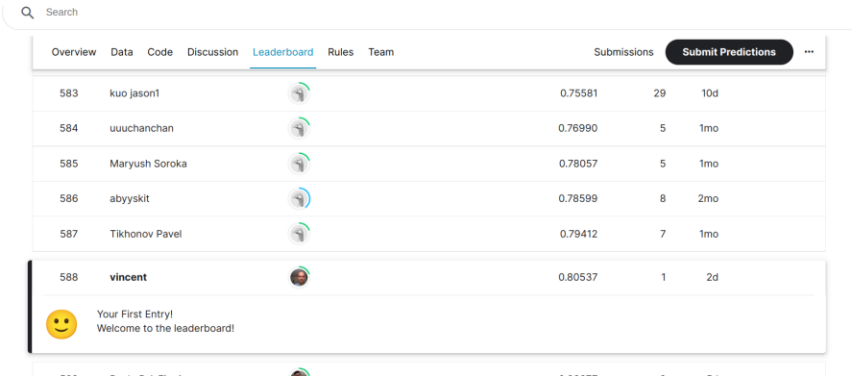
date	id	store_nbr	family	sales	onpromotion	family_cat
2013-01-01	0	1	AUTOMOTIVE	0.0	0	0
2013-01-01	1194	42	CELEBRATION	0.0	0	6
2013-01-01	1193	42	BREAD/BAKERY	0.0	0	5
2013-01-01	1192	42	BOOKS	0.0	0	4
2013-01-01	1191	42	BEVERAGES	0.0	0	3
...	...	...	...	...	...	...
2017-08-15	2999695	25	POULTRY	172.517	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 × 6 columns

Choix du Modèle

- Prophet
  - Librairie avec approche statistique
  - Pas de résultats interessants
- GluontTs
  - Librairie avec approche probabiliste
  - Modèle 'DeepAR' à base de LSTM
  - Donne de bons résultats - out of the box
    - Sans Scaling des données
    - Sans multivariate

Résultats et Suite



- Soumission des résultats
  - Score 0.8
  - 588eme sur +700
  - https://www.kaggle.com/code/vincentjuge/notebookcfae1d1866

- Marges de manoeuvre
  - Data : Ajouter les evenements exceptionnels
  - Modèle : scaler les données
  - Modèle : ajuster les hyper parametres / grid search

Conclusion

- Première compétition !
- Résultats encourageants
  - Participer à d'autres compétitions futures
- Grande Communauté
  - Permet de découvrir de nouvelles approches
  - Par ex : XGBoost est assez utilisé

- Regrets
  - Developpement en local – le kernel kaggle est trop lent / cher
  - Manque de temps pour optimiser
  - On ne peut pas voir les meilleurs notebook pour comprendre et apprendre