Proposition de sujet de thèse

Amélioration des méthodes prévision de ventes pour les biens de consommation

Présentation d'Artefact:

Artefact est une société internationale de services autour de la data, spécialisée dans le conseil en transformation data, dont la mission est de transformer la donnée en délivrant des résultats tangibles sur l'ensemble de la chaîne de valeur des entreprises.
L'approche unique d'Artefact, qui fait le pont entre la donnée et le business, permet à nos clients d'atteindre leurs objectifs business de façon dédiée et efficace. Nos 800 employés allient leurs compétences pluridisciplinaires au profit de l'innovation business des entreprises. Nos technologies de pointe en Intelligence Artificielle, nos méthodes agiles garantissent le succès des projets IA de nos clients, de la conception au déploiement, jusqu'à la formation et l'accompagnement au changement.
Artefact est un cabinet de conseil, avec une forte composante technologique et scientifique.
Depuis plusieurs années années, nous développons des solutions pour la chaine logistiques de grands acteurs du monde de la distribution, tels que Carrefour, Fortenova, L'Oréal ou encore Danone.

Contexte de la thèse & objectifs :

Aujourd'hui, l'immense majorité des chaînes logistiques modernes pour la vente de produits de consommation fonctionnent à flux tendus, pour diminuer les coûts de stockages et les produits invendus.
Aussi, les distributeurs ainsi que les fabricants de biens se doivent de fournir des prévisions opérationnelles de la demande au niveau des produits. Une meilleure précision de ces prévisions ayant des conséquences directes sur les niveaux de stocks, sur la génération de gaspillages, et la satisfaction client.

Ces dernières années, l'enregistrement des données associées à la gestion de la chaîne logistique, aussi bien dans l'e-commerce que dans les "*brick and mortars*", a rendu possible l'utilisation de méthodes d'apprentissage automatique pour aider à une meilleure prévision des ventes.

Historiquement, les problèmes de prévisions de ventes se réalisaient de façon univariée à l'aide de modèles issus du traitement du signal traditionnel, tels que ARIMA ou Exponential Smoothing (et ses dérivées).

Cependant, depuis quelques années, l'apparition de méthodes plus "modernes" de machine learning a eu un impact sur la façon dont on peut modéliser la prévision de ventes, soit avec du gradient Boosting, soit avec des réseaux de neurones profonds comme: Temporal Fusion Transformer, ou DeepAR.

L'objectif de cette thèse est double: À la fois permettre l'amélioration ou la découverte de méthodes statistiques ou de machine learning utilisés par Artefact, tout en s'assurant que ces dernières puissent répondre aux contraintes de systèmes en production de nos clients, sur de grands volumes de données.

A ce titre, Artefact mettra à disposition de la personne doctorante un ensemble de jeux de données réelles issues de systèmes de ventes, couvrant plusieurs années d'histoires, dans différentes industries, avec l'accord de nos clients.

Problématiques scientifiques:

Le signal des ventes qu'on retrouve dans les données dans le monde de la distribution possède un ensemble de spécificités communes :

- Non-stationarité
- Tailles des séries différentes (lancement & fin de produits)
- Effet calendaire (saisonnalités hebdomadaires et mensuelles, effet vacances & jours fériés, etc.),
- Impact des concurrents
- Effets promotionnels
- Effet cannibalisation, halo ou stockage

De plus, les séries temporelles associées possèdent plusieurs caractéristiques hiérarchiques. On dispose généralement d'un arbre hiérarchique naturel entre produits, selon les familles et sous-familles de produits (i.e yaourts aux fruits), marques etc. Dans cet arbre hiérarchique, les produits proches auront des comportements similaires.
Une autre caractéristique hiérarchique est la localisation spatiale des points de vente. Sur certaines gammes de produits, les séries de ventes de deux magasins proches géographiquement ont plus de chance d'être similaires que sur deux magasins éloignés.
Enfin, il faudra être attentif à la gestion des périodes ayant été affectées par des évènements externes comme le COVID ou les gilets jaunes, les ventes ayant été bien fortement affectées par ces phénomènes.

Plusieurs grands axes de développement par rapport aux travaux déjà réalisés par Artefact & la littérature de façon générale. Cette liste des travaux n'est pas exhaustive, et pourra évoluer en fonction des avancées du doctorants et de nos réalisations pratiques chez les clients d'Artefact.

Segmentation du dataset d'entraînement: De façon empirique, nous avons pu observer que parfois les modèles qui obtenaient la meilleure performance étaient entraînés sur un sous-ensemble des séries,

sur la base d'une segmentation hiérarchique. Par exemple, ne prendre en compte que les produits d'une même famille, ou un cluster géographique de point de ventes. De plus, il peut arriver que l'assortiment change au cours du temps, il faudra donc réfléchir à des solutions permettant de prendre en compte ces potentiels changements, garder une cohérence sur le volume total au niveau de la famille, sous famille.
De plus, certains produits qui présentent des saisonnalités annuelles similaires n'appartiennent pas à la même famille de produit, et donc ne profitent pas de la segmentation du dataset. Un des axes de travail sera donc de trouver une méthodologie efficace afin de mieux partitionner le dataset afin de maximiser le transfert d'information entre séries temporelles similaires.

Détection d'anomalies et reconstruction des datasets:
Dans le cas de prévisions pour la chaine logistique, le signal de vente est un bon proxy pour estimer la demande réelle des consommateurs. Toutefois dans le cas de rupture de stocks, la donnée de vente n'est plus représentative de la demande réelle, et il devient nécessaire de reconstruire un historique pour avoir un reflet de la demande théorique.
De la même manière, l'impact de la crise sanitaire du Covid-19, ainsi que les effets des restrictions, a eu un impact sur le signal. Comment retravailler la donnée d'entrée afin de ne pas impacter les prévisions futures en incluant du signal parasite.

Agrégation et désagrégation des séries hiérarchiques: Parfois, l'agrégation de séries temporelles à un niveau supérieur est importante car les séries du niveau inférieur ne possèdent pas assez de signal. Cela peut être le cas pour des séries erratiques, où il n'y a que peu de ventes, ou bien afin de modéliser la cannibalisation entre produits de la même famille.
Un des enjeux des séries hiérarchiques est donc de savoir correctement agréger et désagréger les séries. Un des axes de recherche est donc de trouver de nouvelles méthodes, aussi bien sur l'aspect *bottom-up* que *topdown*, afin de s'assurer une cohérence entre les agrégats et désagrégat.

Séries ultra-saisonnières: Certaines séries temporelles présentent un caractère ultra-saisonnier, c'est-à-dire du signal fort sur une courte période de temps, et l'absence intégrale de signal de vente sur le reste du temps. Quelques exemples emblématiques sont par exemple: les chocolats de pâques, les calendriers de l'avent, les jouets de plages.
Cette ultra saisonnalité entraîne ainsi des challenges pour les méthodes se basant sur des features autorégressives, et il convient de trouver des nouvelles approches pour appréhender ces séries-là.

Politique optimale de ré-entrainement des modèles: Toutes les méthodes développées au cours de cette thèse ont vocation à être déployées dans des infrastructures critiques, sur des très grands volumes de séries temporelles actives. Un des enjeux fort est va être donc de trouver des méthodes optimales pour s'assurer en continuité de la performance des modèles sur l'ensemble des séries, mais aussi de définir des stratégies pour le choix des hyperparamètres des modèles, en dehors de l'approche par force brute. Cet aspect d'autant plus important qu'aujourd'hui, une attention particulière est porté à l'impact carbone que peut avoir un entraînement, et par conséquent on cherche à diminuer au maximum ce temps, tout en conservant des niveaux de performance acceptables.

Références bibliographiques:

Hyndman, Rob J., and George Athanasopoulos. Forecasting: principles and practice. OTexts, 2018.

Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. International Journal of Forecasting 2022

Petropoulos, Fotios, et al. “Forecasting: Theory and Practice.” International Journal of Forecasting, Jan. 2022. Crossref, https://doi.org/10.1016/j.ijforecast.2021.11.001.

Thomassey, Sébastien. "Sales forecasting in apparel and fashion industry: A review." Intelligent fashion forecasting systems: Models and applications (2014): 9-27.

Salinas, David, et al. “DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks.” International Journal of Forecasting, vol. 36, no. 3, July 2020, pp. 1181–91. Crossref, https://doi.org/10.1016/j.ijforecast.2019.07.001.