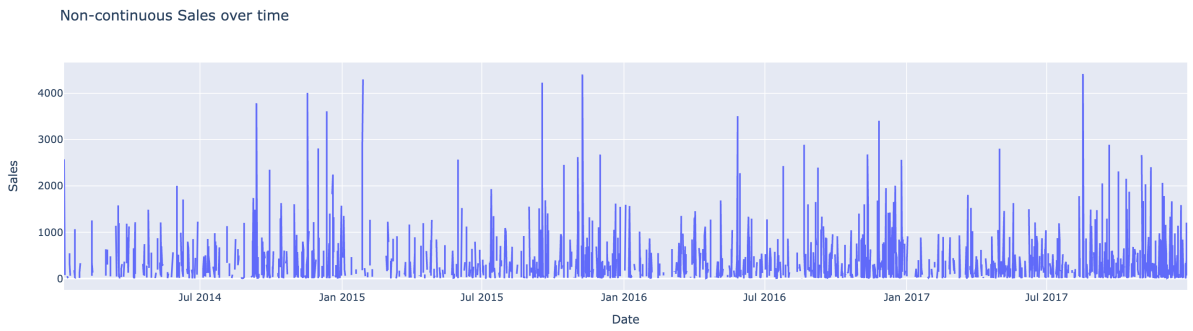
MOOV AI - Sales forecasting

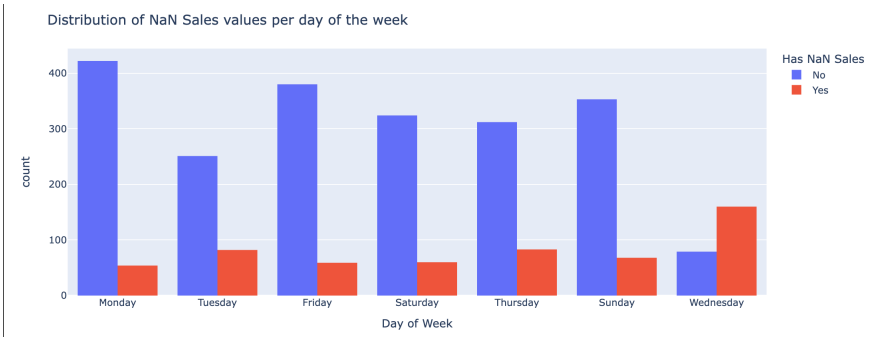
Question #1 : Préparation des données

Enjeux #1 Discontinuité (Intermittence)

Le premier enjeux attrait à la discontinuité dans la série. Sur la période de 2014 à 2017, lorsque la série est à son niveau le moins granulaire, on remarque que 39% des jours sont manquants, ceux-ci s'étalant sur la période en entier.



De plus, c'est période de continuité ne sont pas à interval égale et surviennent de manière aléatoire. Ce qui complique davantage le traitement de la donnée.



Solution:

- Deux features décrivant le temps moyen entres deux ventes l'intermittence: Coefficient of Variation (CV) et l'average demand interval (ADI), et ceci selon son niveau de granularité, pourrait aider l'apprentissage du modèle.
- Remplissage des données avec des 0, l'imputation n'est pas une bonne idée ici, car la valeur est manquante par le simple fait qu'il n'y a pas eu de transaction dans cette journée là.

Enjeux #2 - Granularité au niveau du magasin

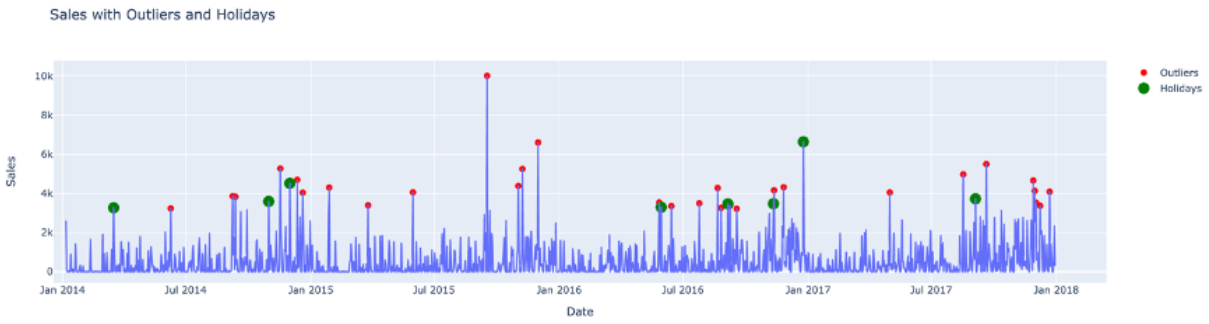
Région	Pourcentage de zéros
Centre	77,1%
Est	73,8%
Sud	83,0%
Ouest	68,3%

Solution:

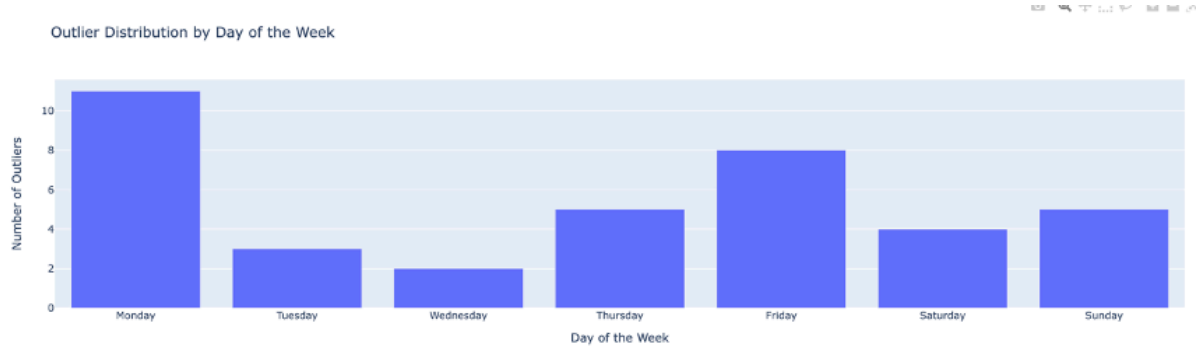
- Il est donc préférable d'agréger les données par semaine, voire par mois, en fonction de l'utilité d'une prévision hebdomadaire/mensuelle pour un responsable et du niveau de granularité d'un magasin.

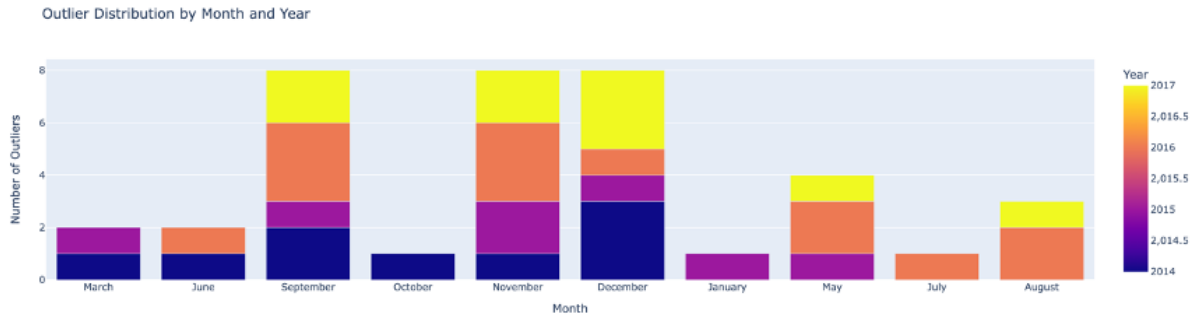
Enjeux #3 - Valeurs sporadiques

Cet enjeux concerne les « valeurs aberrantes » sporadiques, qui sont nombreuses au quotidien et qui ne sont pas totalement aléatoires.



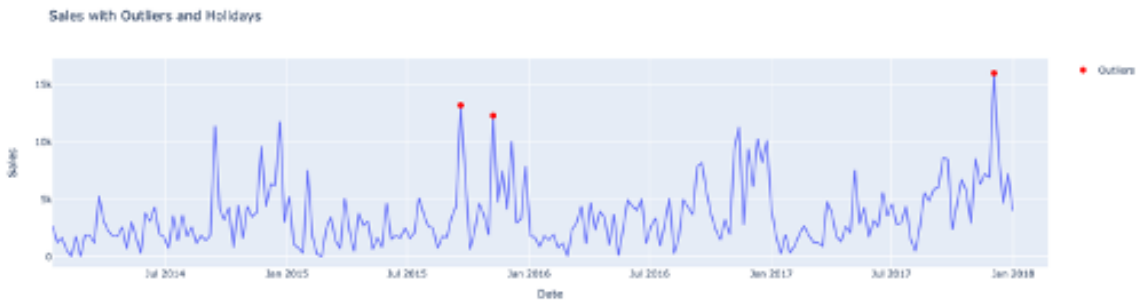
En fait, certaines de ces valeurs aberrantes sont observées lors de vacances nationales ou de grands événements commerciaux tels que le Black Friday. De plus, ces valeurs sporadiques sont plus fréquentes certains jours de la semaine et certains mois.





Solution

- Cela montre l'importance des « caractéristiques » temporelles pour notre modèle, telles qu'une variable catégorielle pour les jours de la semaine et le mois de l'année.
- Cependant, comme nous travaillerons avec des séries hebdomadaires, les enjeux des valeurs sporadiques seront atténués de sorte que le rééchantillonnage « lissera » la distribution et réduira les valeurs sporadiques. Cela facilitera l'apprentissage du modèle.
- De plus, les variables du mois et le fait que la semaine contienne ou non un jour férié national seront également utilisés pour faciliter l'apprentissage.



*Une autre question importante à examiner pourrait être celle des distributions par année. Par exemple, si la distribution en 2014 est significativement différente de celle des années suivantes, nous pourrions être tenté de supprimer l'année 2014 de la formation, ce qui faciliterait l'apprentissage.

Conclusion

La difficulté des données se reflète dans l'intermittence de la série et dans le fait que les directeurs souhaitent recevoir une prédiction au niveau du magasin, ce qui les incite à faire une prédiction granulaire, qui à son tour génère plus de 0. Comme le test ne décrit pas ce qui constitue un magasin, j'émettrai

l'hypothèse qu'un « magasin » peut être décrit par une région et que c'est par région que les ventes futures seront prédites.

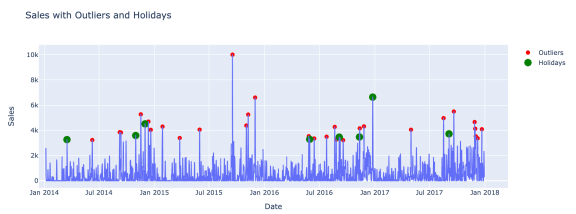
Question #2 : Insights et interprétation

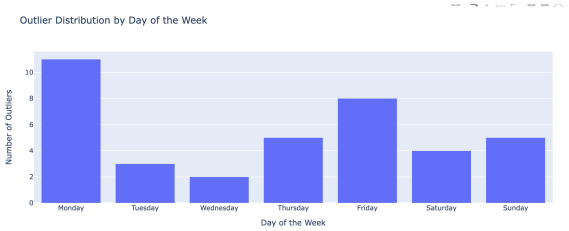
Les insights pouvant permettre aux gestionnaires de mieux comprendre ce qui fait varier leurs ventes sont nombreux.

*L'analyse se fera en partie par région, en supposant qu'une région correspond à peu près à un secteur auquel un cadre est affecté.

L’impact des “holiday”

Tout d'abord, j'étudierai les variables temporelles, telles que la saisonnalité hebdomadaire et annuelle et l'effet des “holiday” sur les ventes.

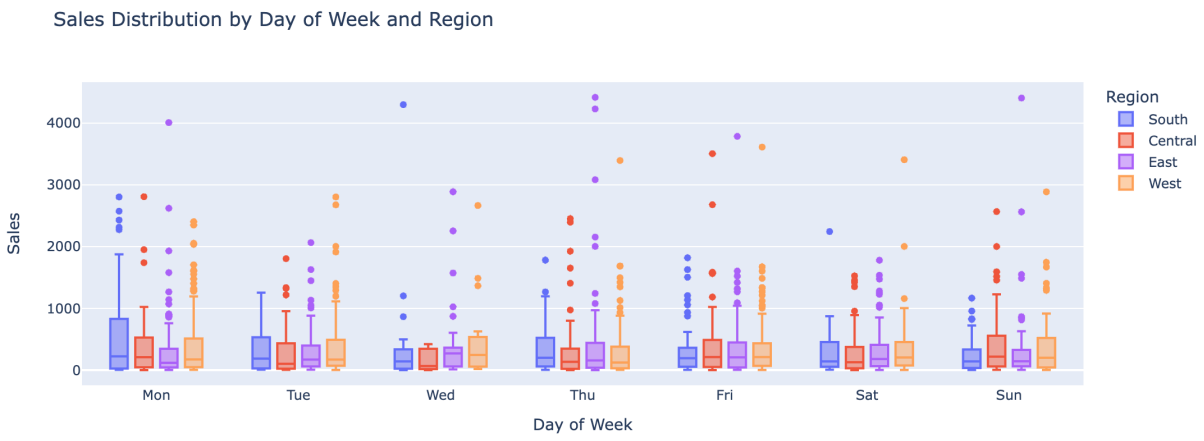




Nous constatons que de nombreuses valeurs aberrantes sont liées à des jours fériés et que ces valeurs aberrantes se produisent principalement les lundis et vendredis.

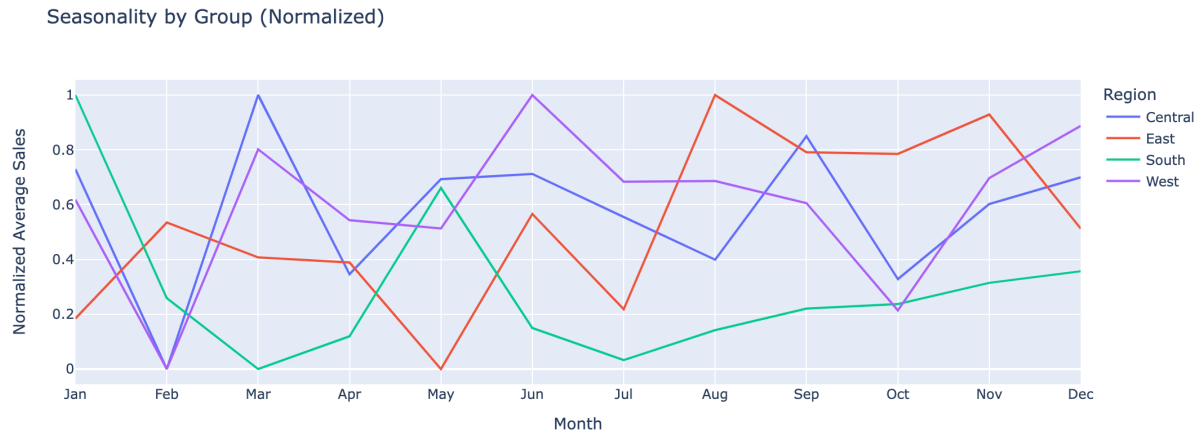
Le client peut donc s'attendre à avoir des stocks plus importants pendant ces jours fériés.

L’impact de la saisonnalité



La saisonnalité au niveau de la semaine montre que dans la région du sud, de plus forte ventes se produisent en moyenne. Les mercredis semblent plus tranquille à travers toute les régions.

Toutefois, il existe une tendance saisonnière assez forte tout au long de l'année, par région.

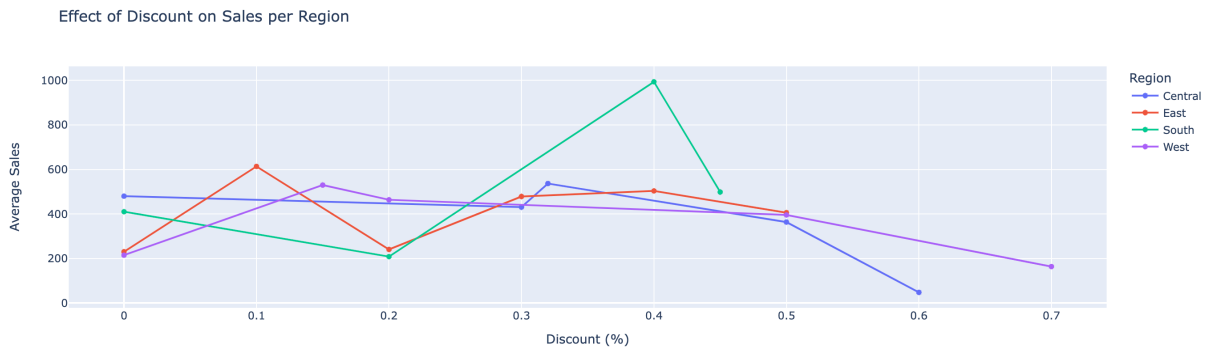


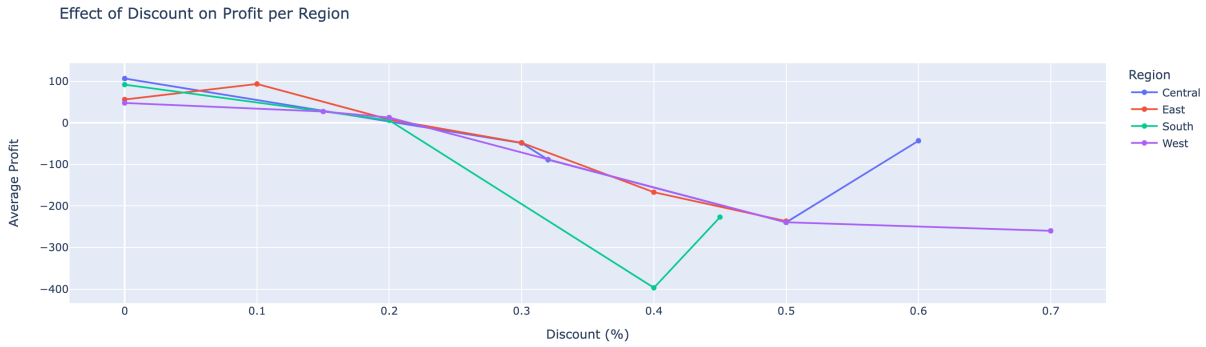
En moyenne, les mois de février, juillet et octobre enregistrent très peu de ventes, tandis que les mois de mars, juin et novembre en enregistrent davantage. Toutefois, nous constatons également que toutes les régions n'ont pas la même saisonnalité et qu'un gestionnaire doit tenir compte du fait que chaque secteur a sa propre dynamique.

En ce qui concerne ma conclusion sur les perspectives temporelles, je pense qu'un gestionnaire peut être conscient que la dynamique des ventes entre ces régions change, et qu'au cours de l'année, ces ventes ne sont pas constantes, ce qui peut lui permettre de stocker davantage en mars, juin et novembre, et moins en février, juillet et octobre. Il en va de même pour les jours fériés.

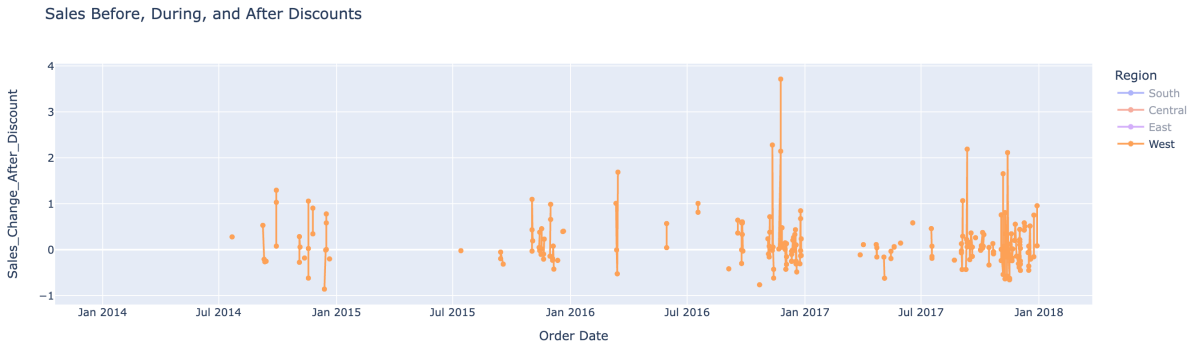
L’impact des rabais

Deuxièmement, je pense qu'il est important d'étudier les bénéfices négatifs, car les ventes ne reflètent pas entièrement ce que les gestionnaires recherchent, c'est-à-dire, en fin de compte, le bénéfice. Ces bénéfices négatifs semblent être dus aux rabais.





Ce que nous remarquons, c'est que des bénéfices négatifs significatifs apparaissent pour des rabais d'environ 0,3 à 0,7, mais la question est la suivante : dans les jours/mois qui suivent ces rabais, les bénéfices augmentent-ils considérablement ?



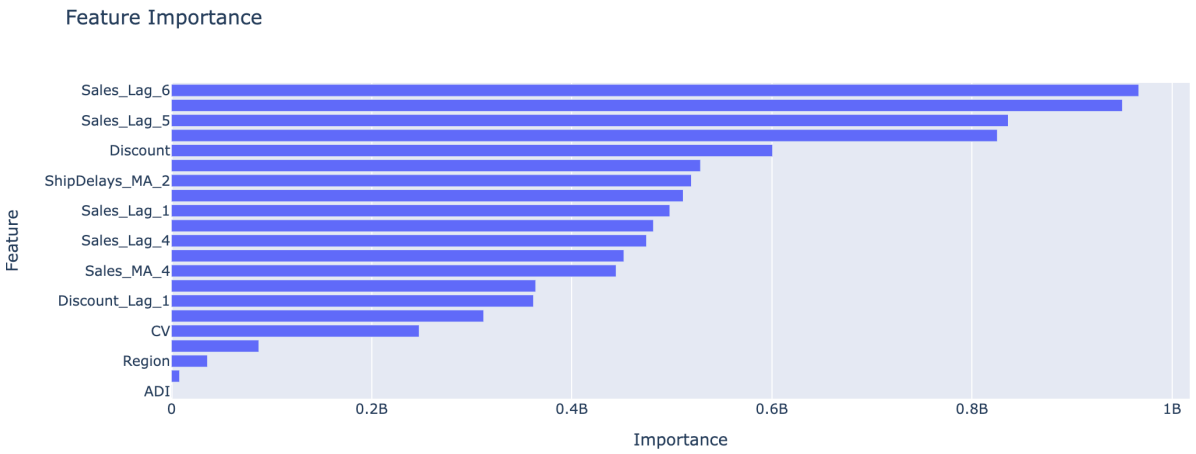
Ce graphique montre l'augmentation ou la diminution des ventes dans les 7 jours suivant une remise. Ce graphique pourrait être étudié plus en détail avec le client afin d'établir la viabilité de ces rabais et le moment opportun pour les mettre en œuvre.

Délais de livraison

Troisièmement, les retards de livraison peuvent entraîner une baisse des ventes, voire même de profit négatifs.

On constate une augmentation significative des ventes avec un délai de livraison de deux jours dans toutes les régions. À ce niveau de granularité, il n'y a pas de schéma clair dans l'impact des délais de livraison sur les ventes ou les bénéfices.

Solution ML



Grâce a un modèle en arbre, nous pouvons évaluer l'importance de certaines caractéristiques, ce qui permet une analyse de ces variables sur les ventes par région au niveau de la semaine.

Il ne faut pas oublier que ces variables sont importantes lorsque l'on fait des prédictions une semaine à la fois, car elles peuvent être différentes en fonction de l'horizon de prédiction. Nous notons ici que les rabais, la semaine et les ventes des semaines passés sont les variables les plus pertinentes pour estimer les ventes futures.

Toutefois, pour permettre au gestionnaire de prendre action sur ces insights, il faudrait davantage étudier l'impact positif ou négatif (augmentation ou réduction des ventes) suite à un mouvement dans ces variables.

Question #3: Solution ML

Prédiction des séries par semaine et par région, en utilisant 2017 comme ensemble de test, et en n'ayant pas d'ensemble de validation, car lightgbm utilise bagging_fraction et bagging_freq, ce qui permet d'utiliser des out of bag observations (OOB) pour sa validation.

Pour prévoir les ventes, j'ai utilisé trois approches, en commençant par une "Seasonal Naive", puis des modèles de séries intermittentes et enfin un modèle d'arbre avec des intervalles de prédiction.

L'utilisation de ces trois modèles est essentielle pour comprendre l'intérêt d'utiliser des modèles plus complexes pour prédire les ventes.

Cependant, malgré une saisonnalité assez forte au niveau de l'année, je n'ai pas utilisé de modèles de séries temporelles classiques tels que l'ARIMA, car j'ai décidé de faire des prédictions par semaine, et à cette agrégation, un grand nombre de 0 étaient présents par région, ce qui peut biaiser les résultats de l'autocorrélation dans l'ACF et l'PACF.

Pourquoi le modèle baseline saisonnier?

Tout simplement parce que nous avons remarqué que les ventes suivent une tendance saisonnière assez forte au fil des ans. L'objectif est d'établir une base de référence simple, mais pas trop naïve. Les inconvénients de cette approche est que le modèle saisonnier de base repose sur l'hypothèse que les ventes d'un mois donné peuvent être approximées par celles du même mois de l'année précédente (exemple : les ventes de janvier 2017 sont prédites à partir des ventes de janvier 2016).

Maintes limitations se modèle possède, tel qu'il **n'utilise aucune variable explicative**, ce qui le rend moins précis face à des changements de conditions

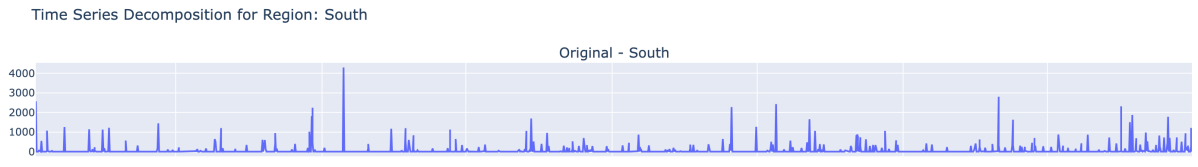
externes.

Pourquoi les modèles de séries intermittentes?

Les modèles intermittents comme **ADIDA (Aggregate-Disaggregate Intermittent Demand Approach)** et **MAPA (Multiple Aggregation Prediction Algorithm)** ne sont pas conçus pour prédire efficacement les observations sporadiques et les périodes avec des observations nulles. Toutefois, ces modèles tentent de prédire intelligemment les séries sachant que les ventes sont irrégulières, comme nous le voyons au niveau journalier ci-dessous pour la région sud.

Ces méthodes se composent de deux parties : l'agrégation et la désagrégation. **ADIDA** utilise une seule agrégation temporelle pour réduire considérablement l'intermittence dans les données. Cette méthode expérimente différentes tailles de répartition pour agréger les données, et le choix de la taille de la répartition peut grandement influencer les performances de prédiction, en raison du fait qu'une classe inappropriée peut entraîner une perte d'informations importantes. Les prévisions sont réalisées pour chaque niveau d'agrégation et sont ensuite agrégées pour générer une prédiction finale à l'indice temporel original. Quant à la désagrégation, cette partie se charge de décomposer la série en fonction de l'indice temporel d'origine. Cette décomposition est effectuée à l'aide de poids, ce qui permet d'assigner des poids plus faibles aux jours où le jeu de données semble comporter moins d'observations, par exemple, les jours de fin de semaine [29].

La méthode **MAPA**, quant à elle, ne se limite pas à une seule taille d'agrégation, mais combine plusieurs tailles d'agrégation. Cette approche permet de capturer différentes tendances et motifs dans la série temporelle, améliorant ainsi la précision des prédictions.



Cette limitation de ne pas capturer les valeurs sporadiques pourrait rendre leur utilisation peu pertinente dans notre contexte d'affaires. Toutefois, si l'objectif du gestionnaire est simplement de maintenir des inventaires non déficitaires pour couvrir sa demande de chaises et de tables, les résultats pourraient être considérés comme intéressants. De plus, ce modèle s'agence bien avec une

mesure de performance adéquate pour évaluer cette perspective : le **Cumulative Forecasting Error (CFE)**.

Un avantage important de ce type de modèle, comparé aux autres modèles, est qu'il pourrait permettre au gestionnaire d'obtenir des prédictions de ses ventes à un niveau plus granulaire pour chaque magasin (ce qui engendrerait plus de zéros).

Modèle LightGBM

Enfin, je pense qu'il est important d'avoir au moins un modèle d'arbre pour permette des intervalles de prédiction, afin de prédire les quantiles 90 et 10 avec un seul modèle. Il faut également disposer d'au moins un modèle utilisant les variables importantes trouvées dans la question 2.

Cependant, ce modèle est plus complexe à mettre en œuvre, et nous devons nous assurer que les intervalles de confiance ont une bonne couverture et sont correctement calibrés.

De plus, ce modèle est plus apte à travailler avec des valeurs nulles qu'un modèle statistique comme ARIMA, et est suffisamment interprétable grâce à la permutation des caractéristiques.

De plus, lightgbm me permet d'appliquer très facilement la régularisation et d'entraîner le modèle en général. L'optimisation a lieu dans l'ensemble d'apprentissage lui-même (OOB), ce qui est à mon avantage car nous n'avons pas assez de données pour un ensemble de validation.

Performance

Un dernier mot sur la performance : ici, le choix du meilleur modèle doit être basé sur la mesure de performance qui est la plus importante pour nous dans le contexte de l'entreprise. Voulons-nous pénaliser les erreurs plus importantes → RMSE ? Ou voulons-nous répartir l'erreur dans le temps → MAE ?

Question #4 : Dégradation de la performance

La dégradation de la performance peut être dû à plusieurs enjeux, par exemple:

1. **Évolution des données** : Les valeurs des variables explicatives ont évolué, reflétant des changements dans l'environnement ou les comportements

des utilisateurs. Cela peut rendre les patterns historiques moins pertinents pour les prédictions actuelles.

2. **Manque de réentraînement fréquent** : Si le modèle n'est pas réentraîné régulièrement, il peut perdre sa capacité à capturer les nouvelles tendances dans les données.
3. **Données insuffisantes dans la nouvelle distribution** : Si la proportion des données reflétant le nouveau comportement est faible, le modèle peut avoir du mal à s'adapter.

Solutions :

1. **Réentraînement régulier du modèle** :
   - Mettre en place une cédule de réentraînement périodique (par exemple, à la semaine ou mensuel) pour ajuster les paramètres du modèle aux données les plus récentes.
   - On pourrait aussi réduire la taille de la fenêtre de données utilisées pour l'entraînement afin de privilégier les observations récentes.
2. **Surveillance des variables clés** :
   - Mettre en place des outils de surveillance des variables explicatives pour détecter les shifts de distribution (par exemple, avec des tests statistiques comme le **K-S test**).
   - Identifier les variables qui ont le plus grand impact sur les prédictions et les monitorer de près.
3. **Validation continue de la performance** :
   - Suivre régulièrement des métriques de performance comme le RMSE ou le MAE sur des ensembles de validation récents pour détecter les baisses de précision dès qu'elles surviennent.

Tout ceci peut se mettre en place à travers des plateformes comme azure Machine learning, cette étape du cycle de développement MLops attrait à la surveillance et la maintenance du modèle.

Question #5 : Intégration de l'IA générative

Une IA générative pour utiliser les informations trouvées, telles que l'effet des vacances ou la saisonnalité par marché et par segment, et pourrait recommander de stocker certains produits spécifiquement à l'avance.

De plus, elle pourrait faire coïncider les offres de réduction avec certaines périodes de l'année où les ventes sont plus faibles.

L'IA générative pourrait aussi travailler à expliquer la prédiction des ventes.

Donne un exemple d'architecture où l'IA générative pourrait être utilisée.

Un modèle de type GPT (comme OpenAI GPT-4 ou Gemini) qui interagit avec les prévisions du modèle ML. Le modèle pourrait:

- Génération de recommandations de stockage avant les périodes de forte demande.
- Proposition de stratégies promotionnelles adaptées aux périodes creuses.
- Explication des prédictions en langage naturel pour les managers.

Pour permettre l'interaction entre le LLM et le modèle ML, un pipeline devrait les connecter. De sorte à ce que les résultats du modèle ML sont envoyés à un LLM (Large Language Model) qui les transforme en recommandations exploitables. De plus, le LLM pourrait également interagir avec des bases de données d'inventaire et de marketing pour proposer des rabais à des moments spécifiques.

Propose un exemple de prompt qui pourrait être utilisé pour interagir avec l'IA générative.

Un bon prompt doit contenir plusieurs éléments permettant au LLM de générer des réponses plus détaillées et précises. Par exemple, il est important de lui donner un **contexte** au problème :

"Tu dois agir en tant qu'assistant pour prévoir les ventes du magasin. Tu reçois des informations du modèle ML qui te permettront de faire des recommandations."

Ensuite, on doit spécifier la **tâche** :

"Tu reçois des estimés des ventes pour les prochaines semaines. Combien de chaises et de tables devrais-tu commander afin de subvenir à la demande

?"

Enfin, on précise les **contraintes** :

"Le délai de livraison des produits est d'environ X semaines."