Note Méthodologique sur la Participation à la Compétition Kaggle "Store Sales - Time Series Forecasting"

Introduction

Cette note méthodologique détaille l'approche adoptée dans ma participation à la compétition Kaggle "Store Sales - Time Series Forecasting". Ce challenge consiste à prédire le nombre de ventes futures de différents magasins en utilisant des données historiques.

L'objectif est de fournir des prévisions précises entre 2017-08-16 et 2017-08-31 de nombre de ventes des produits dans les différents magasins et par famille de produit, toute en précisant les facteurs influençant sur les ventes.

Les données fournies par Kaggle comprennent les ventes journalières de plusieurs produits dans divers magasins, accompagnées d'informations sur les magasins et les produits, des données supplémentaires sont données sur le nombre de transactions journalières par magasin, le cours de pétrole et les jours de vacances locales, régionales et nationales. Ces informations sont fournies sur la période historique et la période à prédire également sauf les transactions qui correspondent aux données historiques uniquement.

Il s'agit d'une problématique de séries temporelles classiques. La complexité de sujet réside dans la façon avec laquelle les données sont présentées par plusieurs catégories (magasin / famille) de qui déroge légèrement à la linéarité des données.

Quatre notebooks ont été développés pour répondre à l'objectif de projet :

1- Notebook d'exploration et de visualisation : comprendre les données fournies, et identifier des tendances et facteurs influentes
2- Notebook de prétraitement : fusionner les différents Datasets en un Dataset global incluant les données d'entrainement et de test
3- Notebook avec un premier modèle xgboost : tester un modèle proposé par l'un des compétiteurs.
4- Notebook avec un deuxième modèle xgboost amélioré : amélioration de modèle précèdent par la création de nouveaux feautures et exploitation de nombre de vente historiques en les réinjectant dans les données d'entrainement.

1. Modélisation et évaluation

Le choix de XGBoost comme modèle principal repose sur sa performance reconnue pour les tâches de régression sur des données temporelles, mais aussi pour contourner la problématique soulignée ci-dessous sur la présentation des données d'entrainement par rapport à une série temporelle classique.

La métrique d'évaluation principale imposée par l'organisateur de la compétition est le RMSLE qui signifie Root Mean Squared Logarithmic Error (Erreur logarithmique quadratique moyenne racine).

C'est une mesure de performance utilisée pour évaluer la précision des prédictions des modèles de régression, en particulier dans les cas où nous voulons minimiser l'impact des erreurs de prédiction disproportionnées pour des valeurs cibles plus grandes

La formule générale pour calculer RMSLE est données par :

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

Ou :

- n est le nombre total d'observations.
- p_i est la prédiction de l'observation i.

- a_i est la valeur réelle de l'observation i.
- *log* est le logarithme naturel.

2. Exploration des données

L'exploration des données fournies a mis en lumière certaines éléments auxquels il a fallu faire très attention :

- Le nombre élevé des magasins et de familles de produits à traiter nous amène à se retrouver avec 33 types de produits multiplié par 54 magasins ce qui donne le nombre total de série par catégorie.
- Les transactions ne peuvent pas être utilisé en tant que features dans les données de test
- Les cours de pétrole présentent des valeurs NAN ou nul pour certains jours historiques
- La table holidays comportent un nombre important de NAN à traiter
- L'analyse de la saisonnalité et la tendance à mis en lumière une tendance croissante et une saisonnalité parfaite liés au jours de vacances et les weekends
- En plus des jours on peut identifier une légère influence du prix de pétrole sur l'évolution générale de des ventes moyennes sur 30 jours.

3. Pré-traitement et préparation de Dataset cible

Suite à l'exploration des données un Dataset global est préparé en traitant les différents points liés au traitement des valeurs manquantes notamment pour identifier les jours de vacances et les jours ordinaires.

En ce qui concerne les prix de pétrole, une technique d'interpolation linéaires est utilisée pour le remplissage des données manquantes.

4. Premier modèle

Un fois le jeu de données global est préparé on split les Dataset en 3 jeux de données :

- Données de test : correspondant à la période que l'on souhaite prédire pour la compétition sachant que sur cette durée aucune données réel n'est fournies, la performance de la prédiction est calculée directement sur Kaggle en présentant les données de prédiction sur la plateforme.
- Données de validation : une partie des données sera réservée pour la validation du modèle. Cette partie est la plus porche chorologiquement aux données de test
- Les données d'entrainement : il s'agit des données historiques sans les donnes sur la durée des données de test et des données de validation.

Une technique de numérisation des données catégoriales est utilisée notamment la date qui a été splittée en trois colonnes pour les jours, les mois et les années. Ensuite une transformation de fourrier est utilisée pour ces 3 variables, les autres variables catégoriels ont été numériser avec des techniques classiques.

Le score obtenu sur kaggle est : 0.94056

5. Amélioration de modèle

Pour améliorer le modèle
Le score obtenu sur kaggle est : 0.94056

6. Limites et piste d'amélioration

La non exploitation des données

Conclusion

La participation à la compétition Kaggle "Store Sales - Time Series Forecasting" a été une opportunité d'appliquer une démarche méthodologique rigoureuse à la prévision des ventes, en utilisant des techniques de visualisation des données, de prétraitement et de modélisation avancées. Malgré certaines limites, l'approche adoptée a permis de dégager des insights pertinents et d'identifier des pistes d'amélioration pour les étapes futures. Les enseignements tirés de cette expérience seront précieux pour les projets futurs de prévision des séries temporelles.