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 de modéliser et prédire le nombre de ventes futures de différents magasins en utilisant des données historiques.

L'objectif est de fournir des prévisions fiables pour la période du 16 au 31 août 2017, en détaillant les ventes par magasin et par famille de produits tout en soulignant les facteurs déterminants les fluctuations des ventes.

Le jeu de données fourni par Kaggle se compose de ventes quotidiennes pour divers produits répartis dans plusieurs enseignes, enrichis par les transactions quotidiennes par magasin, l'évolution des prix du pétrole et le calendrier des jours fériés locaux, régionaux et nationaux. Toutes ces données sont disponibles pour la période des données historiques ainsi que pour la période de prévision, à l'exception des données de transaction qui se limitent à la période historique.

Il s'agit ici d'une question classique de séries temporelles, avec une complexité due à la segmentation des données en multiples catégories (par magasin et par famille de produits), ce qui introduit une légère déviation par rapport au caractère linéaire des données temporelles.

Pour atteindre l'objectif du projet, quatre notebooks ont été élaborés :

1- Notebook d'exploration et de visualisation : pour décrypter le jeu de données fourni, identifier les tendances dominantes et les facteurs d'influence.
2- Notebook de prétraitement : pour consolider les différents jeux de données en un ensemble unifié comprenant à la fois les données d'entraînement et de test.
3- Notebook implémentant un modèle XGBoost initial : pour évaluer une méthode suggérée par un autre participant au concours.
4- Notebook dédié au modèle xgboost optimisé : pour raffiner le modèle initial par le biais de la création de nouvelles variables prédictives et en intégrant les données de ventes historiques au sein du set d'entraînement.

1. Modélisation et Critère d'Évaluation

L'adoption de XGBoost comme modèle principal repose sur sa performance reconnue pour les tâches de régression sur des données temporelles. Cette préférence s'aligne également avec la nécessité de contourner la problématique soulignée ci-dessous sur la présentation des données temporelles.

La compétition exige l'utilisation du RMSLE, (Root Mean Squared Logarithmic Error) ou Erreur Logarithmique Quadratique Moyenne Racine, comme métrique d'évaluation principale. Cette mesure est 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 du RMSLE est la suivante :

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

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

La phase d'exploration des données a révélé plusieurs aspects auxquels il a fallu faire très attention :

- La diversité des magasins et des produits génère une matrice conséquente de séries temporelles, avec 33 types de produits distribués à travers 54 magasins, résultant un total important de séries à catégoriser et analyser.
- L'impossibilité d'intégrer les données transactionnelles comme variables prédictives dans l'ensemble de test, étant donné leur absence pour la période de prévision.
- Des incohérences dans les données sur les prix du pétrole, marquées par des valeurs manquantes (NAN) ou nulles pour certains jours au sein de la période historique analysée.
- La présence de valeurs manquantes dans le jeu de données des jours fériés, nécessitant une stratégie de gestion rigoureuse.
- L'examen des modèles de saisonnalité et de tendance a révélé une corrélation positive entre les périodes de vacances, les weekends et les augmentations de ventes, signalant une saisonnalité et une tendance ascendante marquées.
- Une corrélation légère mais notable entre les fluctuations du prix du pétrole et les tendances générales des ventes moyennes sur une période glissante de 30 jours a également été observée.

Ces constatations initiales ont servi de fondement pour l'élaboration de stratégies de modélisation et de prétraitement des données, assurant la robustesse et la fiabilité des prévisions générées.

3. Nettoyage et préparation du jeu de données

À la suite de l'analyse exploratoire, nous avons constitué un jeu de données consolidé. Le processus de préparation a impliqué des étapes de traitement des valeurs manquantes, particulièrement dans l'identification et la différenciation des jours fériés par rapport aux jours ordinaires, pour garantir l'intégrité des analyses temporelles.

Pour adresser les lacunes identifiées dans les données relatives aux prix du pétrole, nous avons adopté une approche d'interpolation linéaire. Cette méthode permet de combler les valeurs manquantes par des estimations calculées en créant une série continue, en se basant sur les valeurs disponibles avant et après les intervalles vides. Cette technique assure une transition fluide et logique entre les données, préservant ainsi la cohérence temporelle et la pertinence des informations utilisées pour les prédictions.

4. Développement du modèle initial

Après consolidation du jeu de données, nous avons scindé la variable 'date' en trois attributs distincts : jour, mois et année.

Ensuite, nous avons procédé à la sélection des variables prédictives et à la division de jeu de données en trois sous-ensembles distincts :

- Un ensemble de données de test, qui correspond à la période cible de la prédiction pour la compétition. Comme aucune les données réelles ne sont pas fournies pour cette période, l'évaluation de la performance du modèle sur cette période est réalisée directement sur la plateforme Kaggle à travers la soumission des résultats prédits.
- Un ensemble de données de validation, soigneusement sélectionné à partir des données historiques pour être temporellement proche des données de test, ce qui permet d'affiner le modèle avant l'évaluation finale.
- Les données d'entraînement, comprenant l'historique des ventes, sans les périodes définies pour les données de test et de validation.

Un pipeline préconfiguré a été déployé pour la conversion des variables catégorielles, pour l'application de la transformation de fourrier sur les trois variables temporelles (jour, mois et années) et pour la scalarisation des

données. Après entrainement sur le jeu de données d'entrainement, ce pipeline est ensuite appliqué pour la transformation des données de trois sous-ensembles ; entrainement, validation et test afin d'assurer la cohérence.

Après avoir formé un modèle XGBoost en utilisant les données d'entraînement et de validation, le modèle a obtenu un score de performance de 0.971758 sur les données de validation. Lors de la soumission des résultats sur la plateforme Kaggle, le modèle a atteint un score de **0.94056** sur les données de test.

5. Amélioration du modèle

Pour peaufiner notre modèle, nous nous sommes concentrés sur deux aspects principaux :

1. **Enrichissement des Variables Prédictives :** Nous avons intégré des variables catégorielles additionnelles telles que le numéro du magasin et le nom de la ville, ce qui a permis d'ajouter une dimension locale aux prévisions. De plus, 15 nouvelles variables prédictives ont été générées à partir des données historiques de la variable cible, couvrant les périodes allant de 16 à 30 jours avant la date de prévision. Cette approche a pour but de capturer les tendances à court terme et d'offrir une marge suffisante pour anticiper les ventes futures.

2. **Optimisation des Hyperparamètres** : Une recherche approfondie des meilleurs hyperparamètres a été entreprise pour raffiner les performances du modèle. Cette étape cruciale ajuste les paramètres internes du modèle pour maximiser l'efficacité des prédictions.

Le modèle ajusté a démontré une nette amélioration, obtenant un score de performance de 0.777008 sur l'ensemble de données de validation. Après soumission sur la plateforme Kaggle, ce même modèle a réalisé un score de **0.7161** sur l'ensemble de données de test, illustrant sa capacité renforcée à généraliser sur des données non observées.

Conclusion

La participation à la compétition Kaggle "Store Sales - Time Series Forecasting" a été une opportunité d'appliquer une démarche méthodologique rigoureuse pour prédire les ventes, en utilisant des techniques de visualisation des données, de prétraitement et de modélisation avancée. L'approche adoptée a permis d'identifier des facteurs d'influence sur les ventes, de proposer une technique de prédiction des tâches de régression couplée à des techniques d'optimisation des hyperparamètres. D'autres pistes restent à exploiter pour l'amélioration des performances à savoir l'utilisation des modèles spécialisés dans les séries temporelles tel que LSTM (Long Short-Term Memory), et d'autres modèles émergeants tel que les transformer avec PatchTST.