Plan prévisionnel

1. Contexte

Ce travail développe un modèle LightGBM optimisé pour prédire l'affluence quotidienne dans des restaurants japonais, basé sur le data set Recruit Restaurant Visitor Forecasting (Kaggle, 2017). Il agrège réservations Air/HPG, ajoute lags temporels et clusters géo, surpassant XGBoost.

2. Modèle envisagé

LightGBM est l'algorithme retenu pour prédire le nombre de visiteurs dans les restaurants. Il surpasse XGBoost sur ce projet grâce à ses performances supérieures. Il pertmet de predire des valeurs numériques comme le nombre de clients à partir de données historiques.

- Arguments de performance

LightGBM atteint un RMSE de 10.282, MAE de 7.035 et R^2 de 0.612 sur les données de test (11 features : réservations, lags de visiteurs, cluster, etc.). XGBoost est bien plus faible (RMSE 16.332, R^2 0.021 avec 6 features). LightGBM gère mieux les features eng (lags, cluster) et s'optimise via Optuna (meilleur CV RMSE ~10.12).

3. Références bibliographiques

L'état de l'art sur la prévision des visiteurs en restaurant utilise des données historiques, réservations et facteurs externes comme la météo. Le projet s'appuiera sur ces travaux pour enrichir les features et tester plus de modèles.

Review Research Paper: Forecast Restaurant Visitors? Time Series Analysis

Article recherche principal

Shah et al. (2025) comparent XGBoost, Random Forest, LSTM et régression linéaire sur données RFID d'un restaurant finlandais (2019-2024). XGBoost excelle (meilleur MAE/MSE) avec lags, lissage exponentiel, fêtes et météo comme features clés .

Kaggle Recruit (compétition)

La compétition Kaggle Recruit Restaurant Visitor Forecasting (2016-2018) prédit les visiteurs japonais via réservations Air/HPG. Les solutions gagnantes intègrent lags, stats rolling et ML tree-based

4. **Explication de votre démarche de test du nouvel algorithme (votre preuve de concept)**

La démarche commence par une EDA pour nettoyer et enrichir les données, suivie d'un baseline simple pour mesurer les progrès. LightGBM est testé avec optimisation Optuna, et une POC via interface permettra de prédire les visiteurs.

Baseline

Un XGBoost sur les données de test donne RMSE ~16-17 (erreur élevée). Il sert de référence : tout modèle meilleur (comme LightGBM à RMSE 10.28) prouve l'utilité.

Méthode avancée

LightGBM avec 11 features (lags visiteurs, réservations, cluster, etc.), split 80/20 temporel, métriques RMSE/MAE/R^2. Optimisé via Optuna (280 arbres, lr=0.046) et MLflow pour tracking.