### Module 7 Group Proposal

### 

### Team List:

Ashley Sun, Jiaxin Lin, Yumeng Li, Hunter Guo

### Role of Each Member:

Data Preparation/Engineering - Ashley Sun

EDA (High-Level) - Hunter Guo

Feature Engineering - Yumeng Li

Model Selection (Why + Implication) - Jiaxin Lin

### Problem Statement:

Accurate sales forecasting is critical for optimizing inventory management, reducing waste, and ensuring customer demand is met efficiently. This challenge, hosted by the Makridakis Open Forecasting Center (MOFC), aims to improve the precision of daily sales predictions for Walmart stores across three U.S. states: California, Texas, and Wisconsin.

Our team is required to develop robust forecasting models that **predict unit sales at an item level for the next 28 days** using hierarchical sales data. The dataset includes information on store locations, product categories, department segmentation, pricing, promotions, calendar variables, and special events. Traditional time-series models and machine learning techniques are encouraged to enhance predictive accuracy.

The primary goal is to generate **point forecasts** for unit sales, minimizing errors and improving real-world decision-making in retail demand planning. This challenge provides an opportunity to apply cutting-edge forecasting methodologies and advance the field by developing scalable, interpretable, and accurate sales forecasting models.

Success in this challenge will not only contribute to Walmart’s operational efficiency but also refine forecasting techniques applicable across industries, from supply chain logistics to financial planning.

Accurate sales forecasting is critical for optimizing inventory management, reducing waste, and ensuring customer demand is met efficiently. The original dataset provided by the Makridakis Open Forecasting Center (MOFC)includes real information from Walmart on store locations, product categories, department segmentation, pricing, promotions, calendar variables, and special events. We mainly focus on one of the Food Departments in California. Our primary goal is to generate **point forecasts** for department sales revenue, minimizing errors and improving real-world decision-making in retail demand planning at region level.

### Data:

We propose to use the dataset provided by the M5 Forecasting - Accuracy Competition <https://www.kaggle.com/c/m5-forecasting-accuracy/data> . This dataset includes hierarchical daily sales data from Walmart stores in three U.S. states (California, Texas, and Wisconsin), covering various products and departments. Accompanying explanatory variables (prices, promotions, and calendar events) are also available, which makes it a rich dataset for time series forecasting.

### Models:

1. Classical Time Series Models (ARIMA/SARIMA)

We will start with traditional time series methods like ARIMA (AutoRegressive Integrated Moving Average) or SARIMA (Seasonal ARIMA) to capture the autoregressive and seasonal patterns at the product or department level.

1. Tree-Based Ensembles (XGBoost, LightGBM, Random Forest)

Tree-based models are powerful for handling large datasets with many categorical and numeric features. We may engineer features related to lagged prices, store attributes, product categories, promotions, and calendar effects.

Prediction Strategies:

**Top-Down, Bottom-Up, or Middle-Out**: Because the data is structured hierarchically (item, department, store, state, etc.), we will consider various reconciliation methods to ensure consistent forecasts across levels of aggregation

* **Direct Forecasting**: Make separate forecasts for each level of the hierarchy (item, store, department) independently.
* **Bottom-up Forecasting**: First, predict sales at the most granular level (individual items) and then aggregate them up to higher levels (store or department).
* **Top-down Forecasting**: Predict at a higher level (e.g., department level) and distribute the forecast down to the item level based on historical item-to-department ratios.
* **Middle-out Forecasting**: A hybrid approach where you make forecasts for the department level and refine them for specific items.

Model Evaluations:

* **Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** are standard metrics for measuring forecast accuracy.
* **Weighted Root Mean Squared Error (WRMSSE)**: The M5 competition uses WRMSSE as a primary metric, which emphasizes more important products (those with larger sales volumes) and penalizes large errors.

MODEL 1 (ARIMA/SARIMA/SARIMAX):

1. Choose one item to forecast (Bottom-up):
   1. Forecast each state sales, then adding 3 state results together to get national level, compare with actual national sales (pattern catch & MAPE)
   2. Directly forecast at national level, compare with actual national sales (pattern catch & MAPE)

* If (a) did better than (b), bottom-up approach is recommended
* If (a) did worse than (b), bottom-up approach is not recommended

1. Choose one item to forecast (Top-down):
   1. Forecast at national level, then times each state percentage to get state level forecast result, compare with actual state sales (pattern catch & MAPE)
   2. Directly forecast that state sales, compare with actual state sales (pattern catch & MAPE)

* If (a) did better than (b), top-down approach is recommended
* If (a) did worse than (b), top-down approach is not recommended

1. If one item\_id has sparse data (0 sales for some days) at daily level, use department\_id at daily level (department contains multiple items, ensuring sales volume > 0 each day):
   1. For exogenous variables (events), it happens on daily level
   2. ML model needs granular data, it is better to do at daily level than weekly level

MODEL 2 (ML):

1. When using ML models, we don’t model data in time series fashion, we create time-based features (use one-hot encoding to transform into binary time features)
2. Depending on features, we might also add SD, min/max, coefficient of variation as data features
3. ML (GBRT & XGBoost) models usually do better at state level (Bottom-up is better)
   * see if our final results can validate this

MODEL 3 (DL - Neural Network Model) (OPTIONAL):

Same as above

Updates:

* Choose a department\_id (Foods3) and forecast that department’s revenue (units sold\*price, make business sense)
* Choose a state (CA) and don’t aggregate to national level (since only 3 states)
  + Can do store to state (bottom-up)? Then state to each store (top-down)?
  + Can do daily to weekly to monthly to yearly (bottom-up)? Vice versa (top-down)?
* Data from 2011/1 - 2016/6 (Total 5.5 Years = 1969 Days)
* Choose major events (exogenous variable), since many events happen less in a year (Religious 56 > National 52 > Cultural 41 > Sporting 18 of 1969 Days)
* SNAP purchase in CA (65% False, 35% True)
* Bottom-up OR Top-down?

Foods3\_CA\_df: agg + 4 stores time revenue’s time series data 2011-2016 + 5 major events (Christmas, Thanksgiving, New Year, Labor Day, Super Bowl)

最后记得改SAP邮件