Data Wrangling for Food Demand Forecasting

Data Acquisition

The data was provided by the meal delivery company. The dataset, present in csv format, contains

1. Product(Meal) features such as category, sub-category, current price and discount.
2. Information for fulfillment center like center area, city information etc.
3. Historical data of demand for a product-center combination (Weeks: 1 to 145)
4. Future data of product-center combination for prediction (Weeks: 146 to 155)

Data Analysis

Below steps were taken to analyse the dataset

Data	Analysis Steps	Conclusion
Meal Information	● Missing Values ● Data Types of each column ● Distribution of category and cuisine columns	There is nothing to clean.
Fulfilment Center Information	● Missing Values ● Data Types of each column ● Distribution of city_code, region_code and center_type columns ● Statistics of op_area column like mean, percentile, etc	There is nothing to clean.
Training data / Historical data	● Missing Values ● Data Types of each column ● Distribution of week, center_id, meal_id, emailer_for_promotion and homepage_featured columns ● Statistics of num_orders, checkout_price and base_price columns	● Records are missing for some week, center and meal combination ● Data contains two outliers, one with 24299 as number of orders and another as 2.97 checkout_price.

Test data / Future data	• Missing Values • Data Types of each column • Distribution of week, center_id, meal_id, emailer_for_promotion and homepage_featured columns • Statistics of num_orders, checkout_price and base_price columns	Records are missing for some weeks, center and meal combination.

Outliers

Action on outliers will be taken during the modeling based on the performance of model with and without outliers.

Missing Records

Missing records can be because to below two reasons:

1. There is actually no sales for that meal, center and weeks combination
2. Records were not captured due to technical error

Reason will become more clear after exploring data and then action can be taken.

Data Merging

All three data are present in different dataframes. Hence, its required to merge them into one dataframe. Below steps were taken to merge the dataset

1. Left join on training data and meal information on meal_id.
2. Left join on training data and fulfilment center information on center_id.

Same steps were taken for test data.

Derive new variables

Below are the new variables, related to number of orders, derived from existing dataset

Variable Name	Description	Derived from
average_orders_Nweek	It is the mean of num_orders for particular meal_id and center_id in past few weeks. N -> 13, 26 and 52	• center_id • meal_id • week • num_orders

average_orders_Nweek_across	It is the mean of num_orders for particular meal_id across all centers in the past few weeks. N -> 13, 26 and 52	• meal_id • week • num_orders
average_orders_Nweek_adj	It is the mean of num_orders for particular meal_id and center_id in past few weeks ending at 10 weeks in the past. e.g:- for week 50, past weeks will be 37-40 weeks. N -> 13 and 26	• center_id • meal_id • week • num_orders
average_orders_Nweek_adj_across	It is the mean of num_orders for particular meal_id across all centers in the past few weeks ending at 10 weeks in the past. N -> 13 and 26	• meal_id • week • num_orders

Below are the new variables, related to time periods, derived from week column

Variable	Description
year	It represents the year, group of 52 consecutive weeks, in which the record belongs.
month	It represents the month, group of 4 consecutive weeks in a year, in which the record belongs. Since, month is considered as a set of 4 weeks, there are 13 months in the dataset.
quarter	It represents the quarter, group of 13 consecutive weeks in a year, in which the record belongs.
week_in_month	Since, month contains set of 4 weeks, this variable represents record belongs to which of these 4 weeks.

Below are the new variables related to prices of meals.

Variable	Description	Derived from
mean_base_price	It is the mean of all base_price for a particular center_id and meal_id till that week	● center_id ● meal_id ● week (<= current record) ● base_price
discount	It is the discount (in percentage) that customers got in that week for a meal in that center.	● mean_base_price ● checkout_price