Milestone Report

Food Demand Forecasting Challenge

Monil Gudhka

19th September, 2019

Contest: https://datahack.analyticsvidhya.com/contest/genpact-machine-learning-hackathon-1/

Repository: https://github.com/monilgudhka/food_demand_forecasting

Problem Statement

Demand forecasting is a key component to every growing online business. Without proper demand forecasting processes in place, it can be nearly impossible to have the right amount of stock on hand at any given time. A food delivery service has to deal with a lot of perishable raw materials which makes it all the more important for such a company to accurately forecast daily and weekly demand.

Too much inventory in the warehouse means more risk of wastage, and not enough could lead to out-of-stocks — and push customers to seek solutions from your competitors.

The client is a meal delivery company which operates in multiple cities. They have various fulfillment centers in these cities for dispatching meal orders to their customers. The client wants to forecast the demand in these centers for upcoming weeks so that these centers can plan the stock of raw materials accordingly.

The replenishment of majority of raw materials is done on a weekly basis and since the raw material is perishable, the procurement planning is of utmost importance. Secondly, staffing of the centers is also one area wherein accurate demand forecasts are really helpful.

The evaluation metric for this competition is 100*RMSLE where RMSLE is Root of Mean Squared Logarithmic Error across all entries in the test set. Since, we do not have access to the output of test set. Hence, Partial evaluation will be done on validation set and final evaluation will be done by submitting the solution in the Contest.

Dataset

The client has provided the following information, the task is to predict the demand for the next 10 weeks (Weeks: 146-155) for the center-meal combinations in the test set:

1. **Weekly Demand data (train.csv)**: Contains the historical demand data for all centers, test.csv contains all the following features except the target variable

Variable	Definition
id	Unique ID
week	Week No
center_id	Unique ID for fulfillment center
meal_id	Unique ID for Meal
checkout_price	Final price including discount, taxes & delivery charges
base_price	Base price of the meal
emailer_for_promotion	E-Mailer sent for promotion of meal
homepage_featured	Meal featured at homepage
num_orders	(Target) Orders Count

2. **fulfilment_center_info.csv**: Contains information for each fulfilment center

Variable	Definition
center_id	Unique ID for fulfillment center
city_code	Unique code for city
region_code	Unique code for region
center_type	Anonymized center type
op_area	Area of operation (in km^2)

3. **meal_info.csv**: Contains information for each meal being served

Variable	Definition
meal_id	Unique ID for the meal
category	Type of meal (beverages/snacks/soups….)
cuisine	Meal cuisine (Indian/Italian/…)

Data Cleaning

After analysing the dataset, two issues were found.

Outliers

Data contains two outliers,

1. Record with 24299 number of orders
2. Record with 2.97 checkout_price

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

Missing Records

Records are missing for some weeks, center and meal combination. These can be because of following reasons

1. There is actually no sales for that meal, center and weeks combination
2. Center does not take orders of that meals
3. Records were not captured due to technical error

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 (Feature Engineering)

After Merging the data into a single dataset, we derive new variables using existing variables and past records.
Deriving new variables based on the past number of orders.

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

Deriving new variables by grouping consecutive weeks into one parent class.

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.

Deriving new variables from the past base price and checkout price of meals.

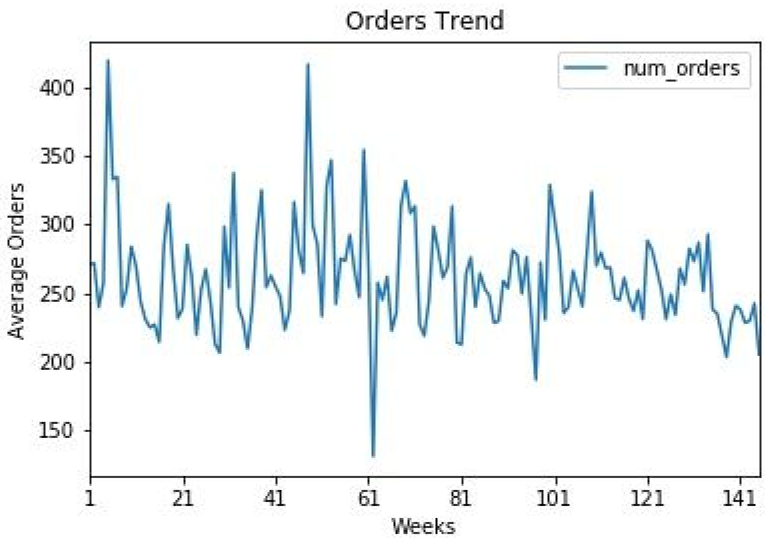
Variable	**Description**	**Derived from**
mean_base_price	It is the mean of all base_price for a particular center_id and meal_id till that week	● center_id ● meal_id ● week (<= current record) ● base_price
discount	It is the discount (in percentage) that customers got in that week for a meal in that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

Exploratory Data Analysis

Initial Questions

1. What is the trend in number of orders irrespective of center and meal?
2. What is the trend in number of orders with respect to meal?
3. What is the trend in number of orders with respect to center?
4. What is the impact of promotional activities like email and homepage plays on number of orders?
5. Are there any correlation between checkout_price, base_price with number of orders?

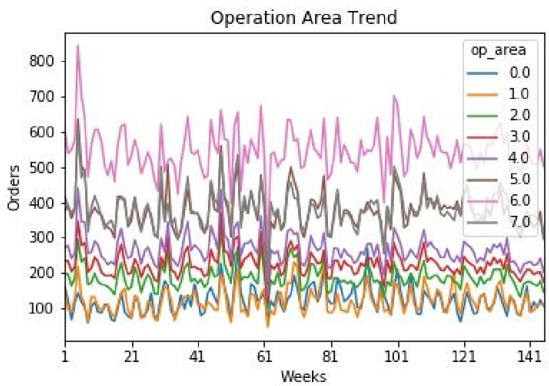
Overall Orders Trend

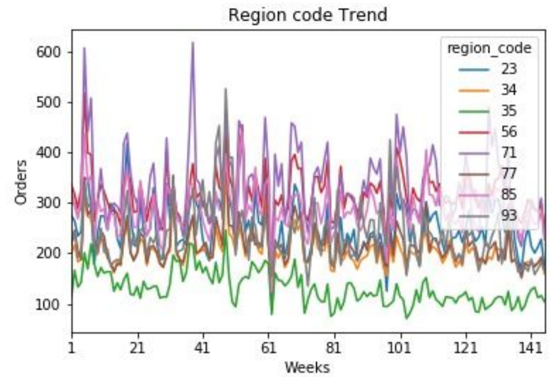


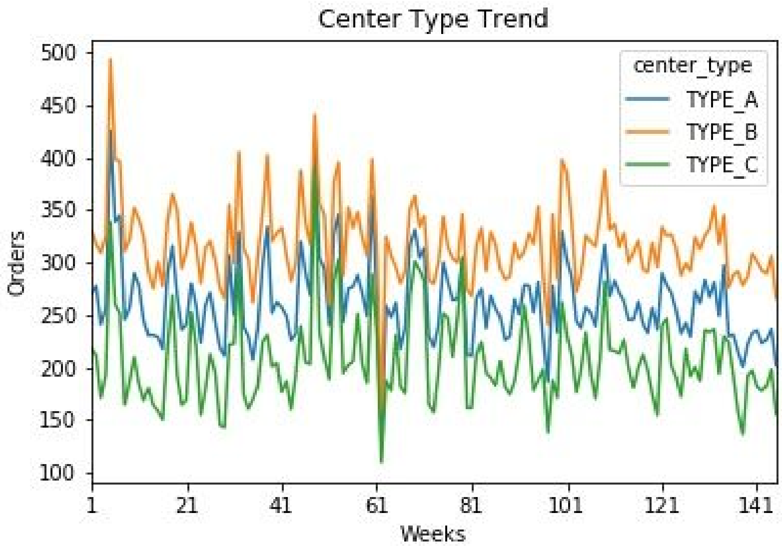
Above plots represents the Monthly, Week in month and weekly Orders Trend respectively. Below are findings from above plots:

1. It was found that week 62 had lowest orders while week 5 and week 48 had highest orders.
    a. After further analysis, there was hugh difference in the promotional activity by emails for week 62 compared to week 48 and week 5.
2. It was found that month 2 had highest orders and month 9 had the lowest orders.
3. It was found that start and end of the month has highest orders as compared to the mid of month.
4. Data is not sufficient to analyse the yearly trend in number of orders.

Center Wise Orders Trend



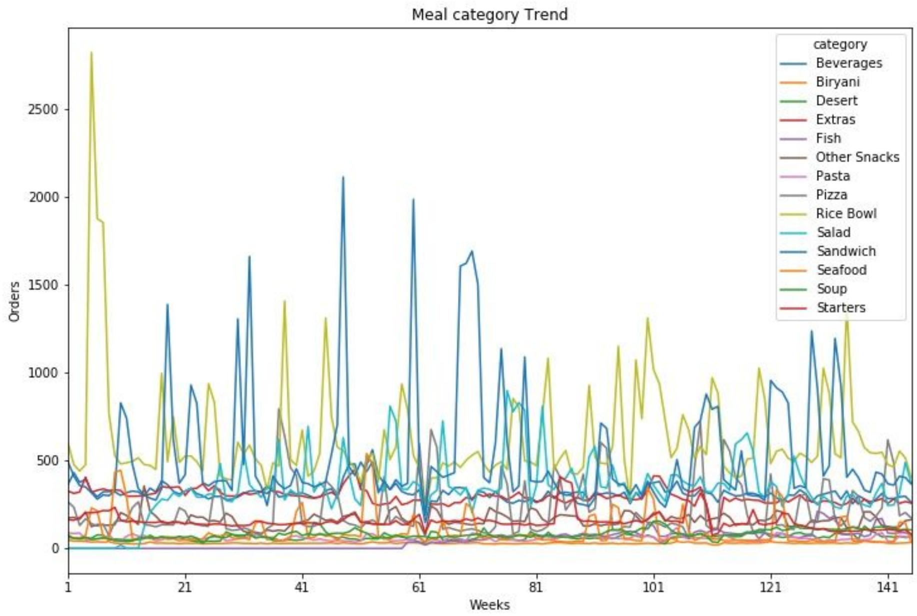


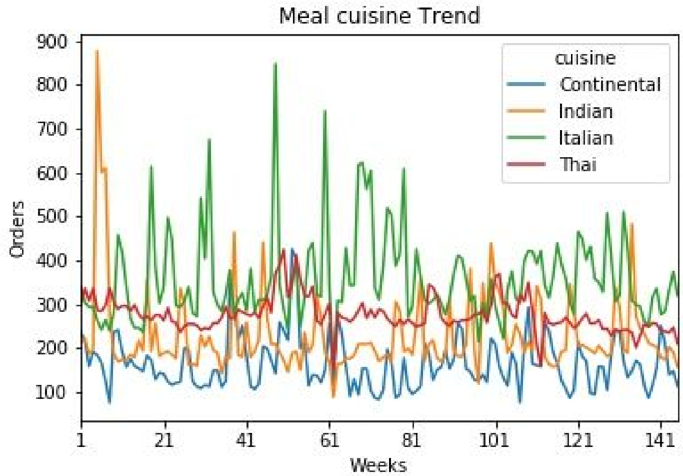


Above plots represents the weekly order trend with respect to the center's operation area, region code and center type respectively. Below are findings from above plots:

1. Centers with center type TYPE_B get more orders than centers with center type TYPE_A and TYPE_C
2. Orders increased with increase in operating areas
3. Centers with region code 35 has lowest orders
4. There are fluctuations in the number of orders for almost all regions and hence, cannot contribute to the problem statement

Meal wise Orders Trend





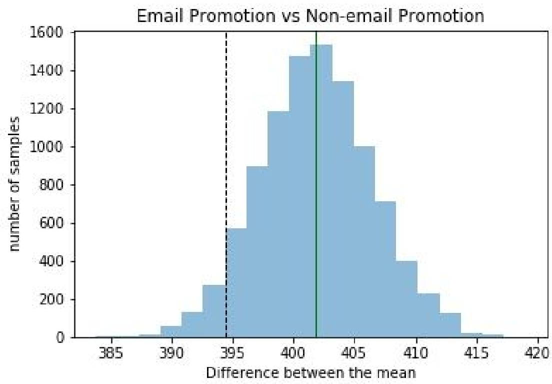
First plot represents the weekly orders trend in meal category and second plot represents the same in cuisine. Below are findings from both the plots:

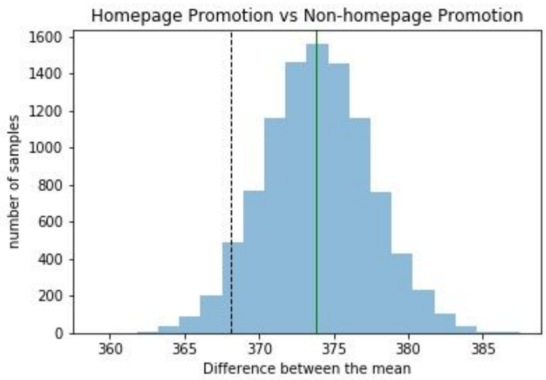
1. Orders for Italian meals and Beverages are always high
2. Orders for Salad increased after week 18
3. There are fluctuations in the number of orders for Indian meals, Rice Bowl and Sandwich

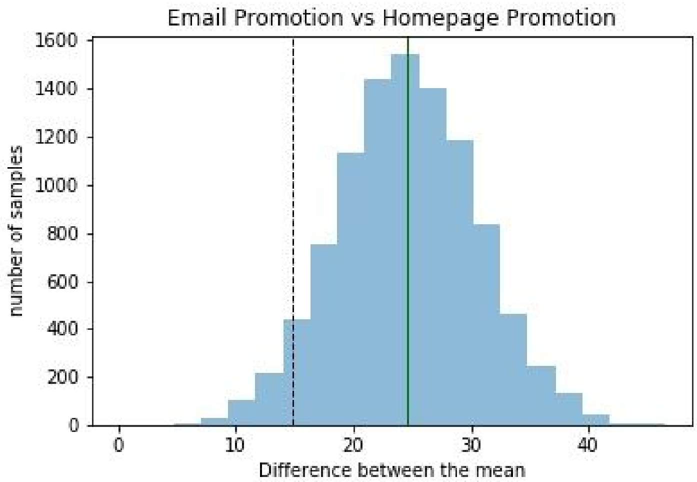
Promotional Activities

Below are the questions to identify the impact of promotional activity on number of orders

1. Does promotion by email results in increase in number of orders?
2. Does promotion in homepage results in increase in number of orders?
3. Since, there can be activity in any one way, which promotional activity has higher impact on number of orders?



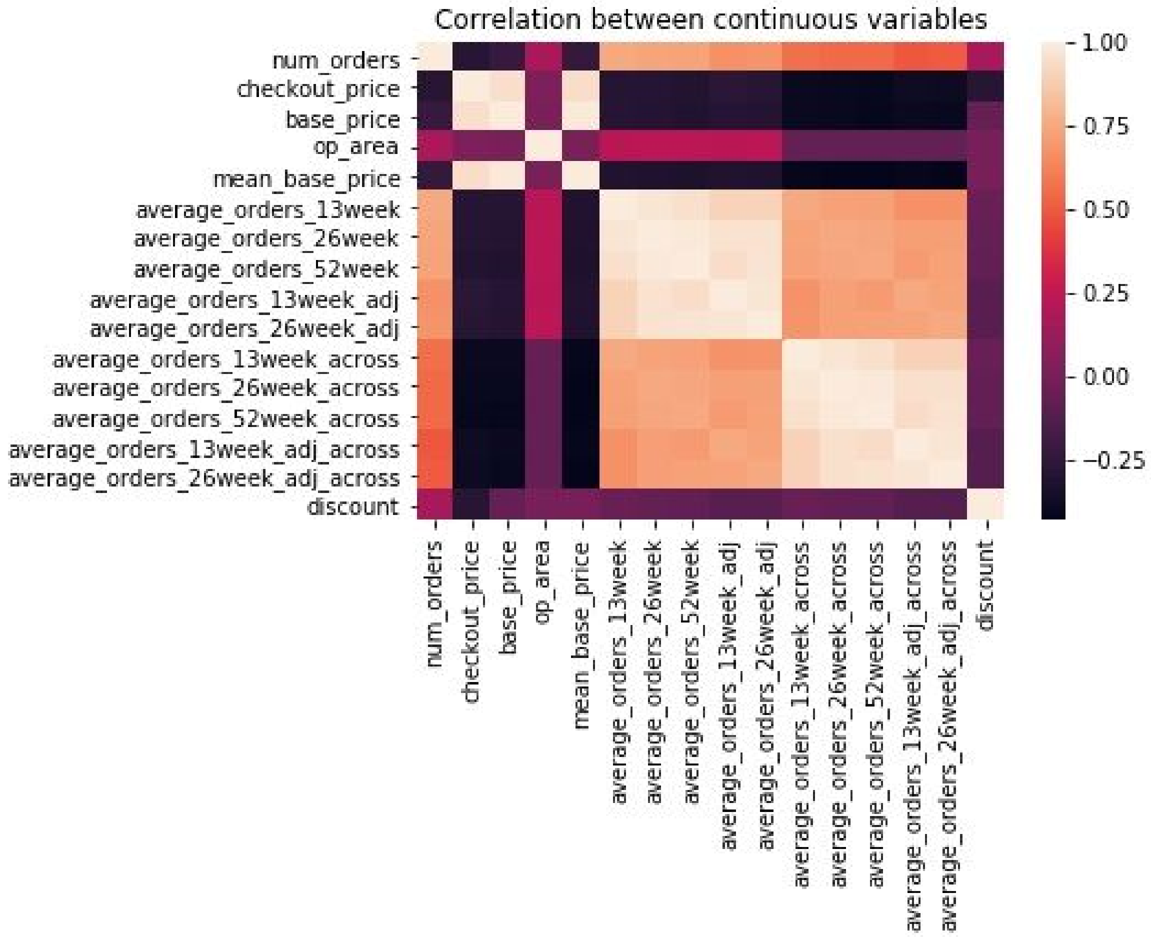




All the above questions were answered using the hypothesis test. The distribution of difference in mean of number or orders, one distribution for one question, are displayed above. Below are outcome of tests:

1. Promotion Activity by emails increases the number of orders
2. Promotion Activity in homepage also increase the number of orders
3. Promotion Activity in homepage has more impact than emails on increase in number of orders

Correlation between price and number of orders



Above heatmap displays the correlation between all the continuous variables present in the dataset. Below are some findings after analysing above heatmap:

1. The checkout price and base price have high positive correlation with each other
2. Both prices also have negative correlation with number of orders
3. Since, mean base price is derived from base price of past orders. Hence, it have the same correlation as that of base price with other variables
4. Discount, which was derived from checkout price and mean base price, have low positive correlation with number of orders
5. Discount have low negative correlation with checkout price