Food Demand Forecasting Challenge

A report by Monil Gudhka

About

Title:
Food Demand Forecasting Challenge

Contest:
[Genpact Machine Learning Hackathon]

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

Problem Statement

- Client: meal delivery company
- Problem:
  - Deals with a lot of perishable raw materials
  - Not enough inventory -> out-of-stocks -> push customers to competitors
  - Too much inventory -> more risk of wastage
- Solution will also help in
  - Planning the stock of raw materials
  - Staffing of the centers

Evaluation Metric

100 * Root of Mean Squared Logarithmic Error (RMSLE)

across all entries in the test set

Dataset

	id	week	center_id	meal_id	checkout_price	base_price	emailer_for_promotion	homepage_featured	num_orders
0	1379560	1	55	1885	136.83	152.29	0	0	177
1	1466964	1	55	1993	136.83	135.83	0	0	270
2	1346989	1	55	2539	134.86	135.86	0	0	189
3	1338232	1	55	2139	339.50	437.53	0	0	54
4	1448490	1	55	2631	243.50	242.50	0	0	40

	center_id	city_code	region_code	center_type	op_area
0	11	679	56	TYPE_A	3.7
1	13	590	56	TYPE_B	6.7
2	124	590	56	TYPE_C	4.0
3	66	648	34	TYPE_A	4.1
4	94	632	34	TYPE_C	3.6

	meal_id	category	cuisine
0	1885	Beverages	Thai
1	1993	Beverages	Thai
2	2539	Beverages	Thai
3	1248	Beverages	Indian
4	2631	Beverages	Indian

Outliers & Missing Records

- Outliers:
  - Record with 24299 number of orders
  - Record with 2.97 checkout_price
- Missing records:
  - No Orders of some product-center combination for some week

Merge Data

	id	week	center_id	meal_id	checkout_price	base_price	emailer_for_promotion	homepage_featured	city_code	region_code	center_type	op_area	category	cuisine
0	1379560	1	55	1885	136.83	152.29	0	0	647	56	TYPE_C	2.0	Beverages	Thai
1	1466964	1	55	1993	136.83	135.83	0	0	647	56	TYPE_C	2.0	Beverages	Thai
2	1346989	1	55	2539	134.86	135.86	0	0	647	56	TYPE_C	2.0	Beverages	Thai
3	1338232	1	55	2139	339.50	437.53	0	0	647	56	TYPE_C	2.0	Beverages	Indian
4	1448490	1	55	2631	243.50	242.50	0	0	647	56	TYPE_C	2.0	Beverages	Indian

Feature Extraction

Based on past orders

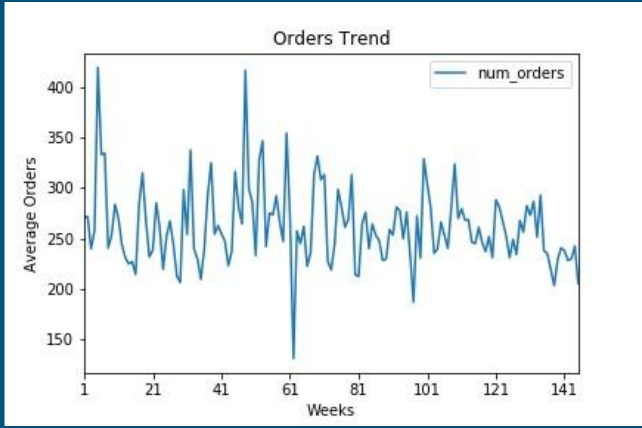
- average_orders_Nweek
- average_orders_Nweek_across
- average_orders_Nweek_adj
- average_orders_Nweek_adj_across
- mean_base_price
- discount

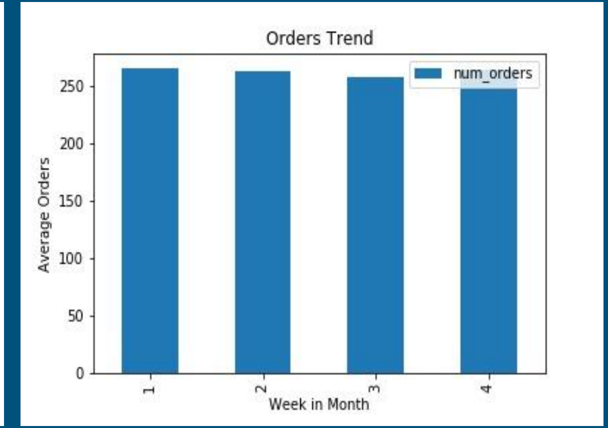
where N is 13, 26 and 52

Based on weeks

- year
- month
- quarter
- week_in_month

Analysis: Overall Orders Trend



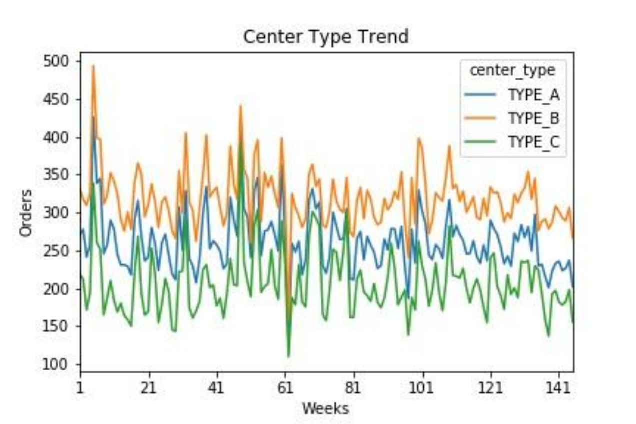


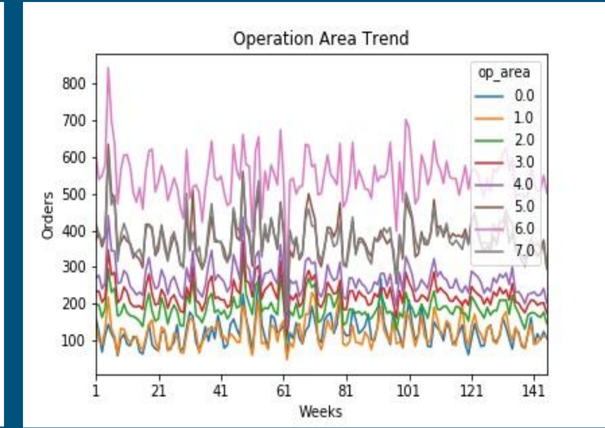
- Week 62 have lowest Orders
- Week 5 and 48 have highest Orders
- Because of Promotions by emails

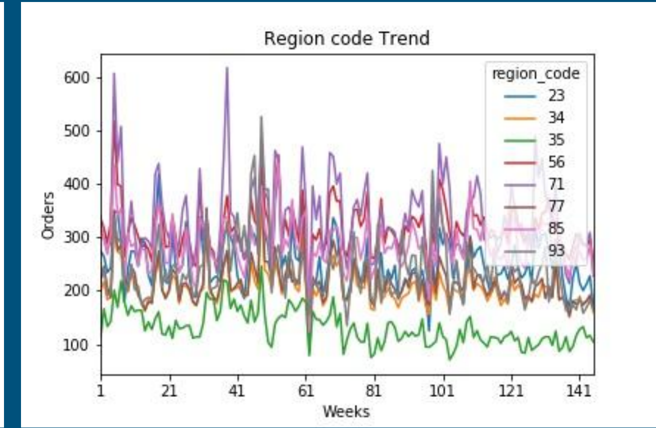
Start and end of the month has highest orders

- Month 2 have highest orders
- Month 9 have lowest orders

Analysis: Center Wise Orders Trend







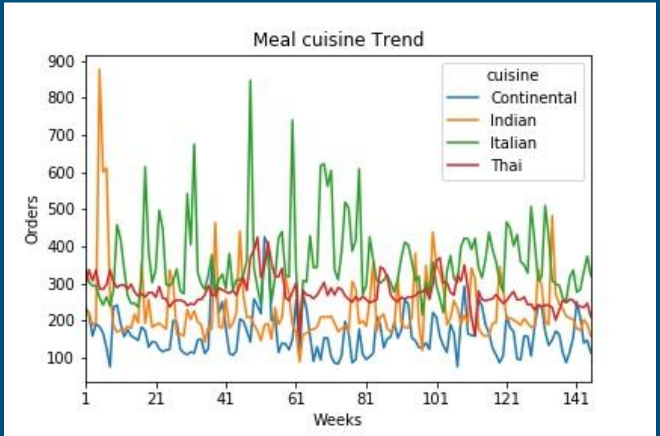
- TYPE_B has highest orders
- TYPE_C has lowest orders

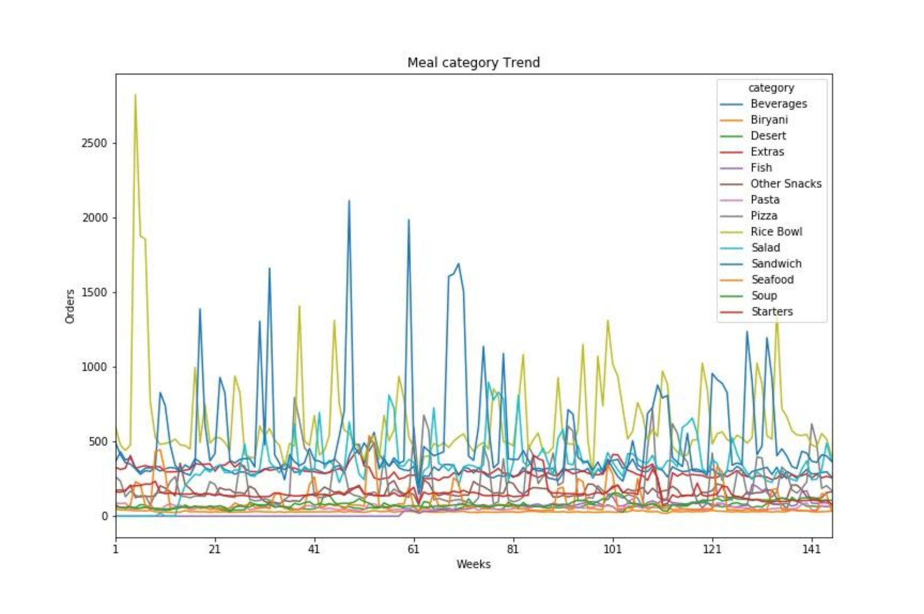
Positive correlation between Operation Area and Orders

- Region code 35 have lowest orders
- Fluctuations for almost all regions

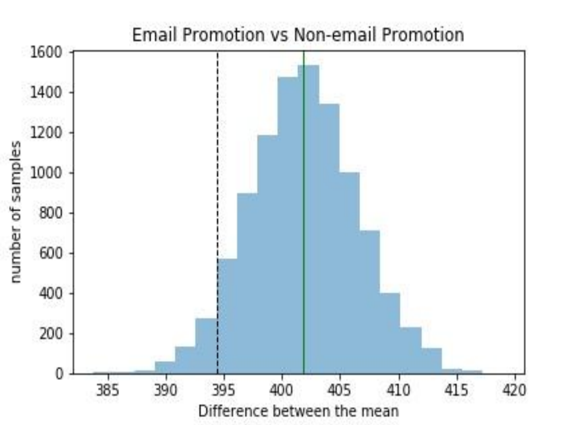
Analysis: Meal wise Orders Trend

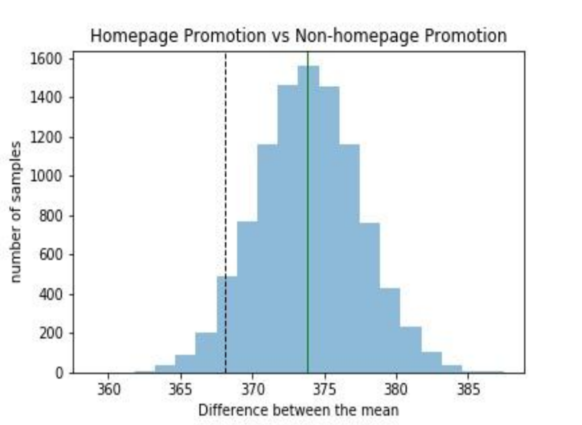
- Italian meals and Beverages has high Orders
- Orders for Salad increased after week 18
- Fluctuations for Indian meals, Rice Bowl and Sandwich

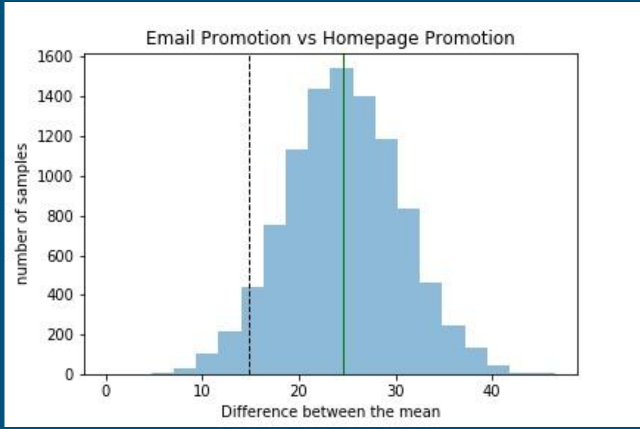




Analysis: Promotional Activity







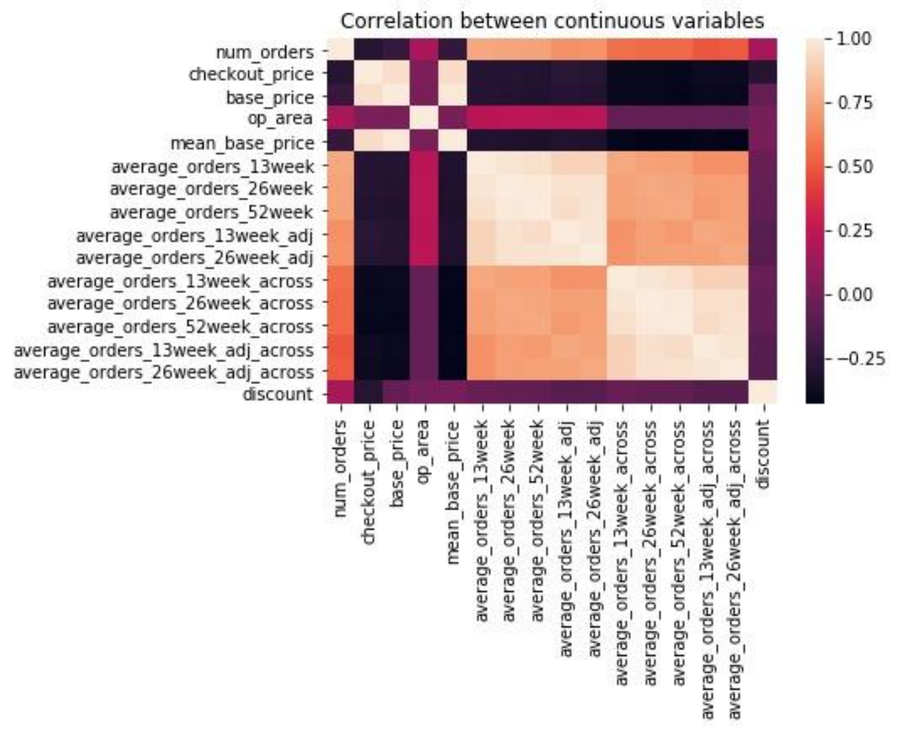
Promotions by emails increases the number of orders

Promotions in homepage increases the number of orders

Promotions in homepage has more impact than emails

Analysis: Price

- High positive correlation between checkout price and base price
- Negative correlation between Orders and both prices
- Low positive correlation between discount and Orders
- Low negative correlation between discount and checkout price



Ineffective Features

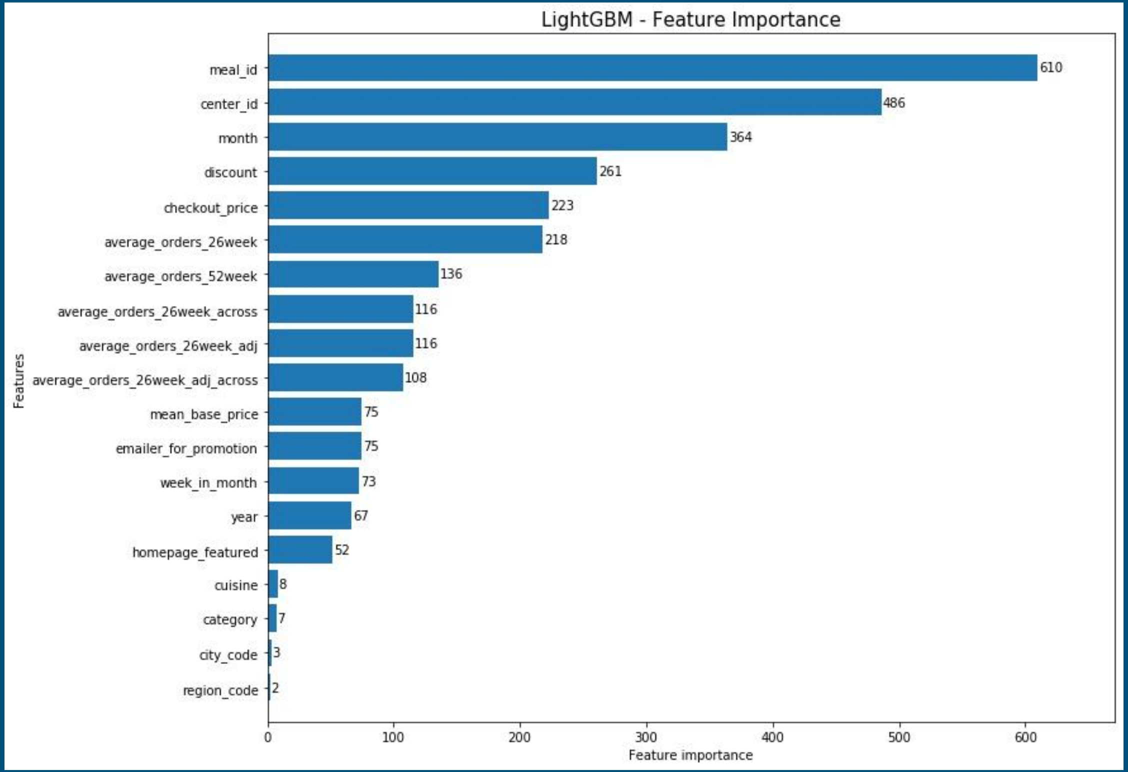
Based on Analysis

- base_price
  - mean_base_price is the better
- quarter
  - month is more granular
- average_orders_13week
- average_orders_13week_across
  - Information not available most of time
- week
  - Train set: 1-145
  - Test set: 146-155

Based on modelling

- average_orders_13week_adj
- average_orders_52week_across
- average_orders_13week_adj_across
  - reduces performance
- op_area
  - Algorithm finds it redundant

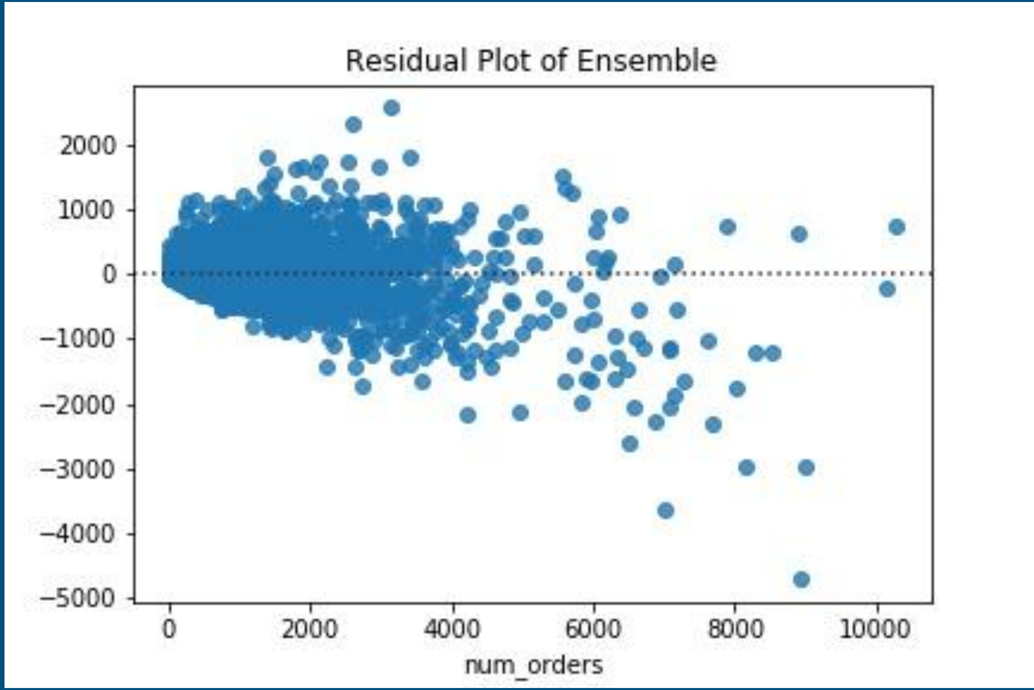
Feature Importance



Modelling

With Outliers	51.0826	>	51.3646	Without Outliers
One Hot Encoding	51.0484	>	51.0826	Label Encoding
Raw values	51.0826	>	51.4003	Natural Logarithm
Tuned LightGBM	50.5356	>	50.5686	Tuned XGBoost
<u>Combining models</u>	<u>50.2260</u>	>	50.5356	Individual model

Residual Plot



Further Improvements

1. More features related to the centers
2. Algorithms other than LightGBM and XGBoost
3. Parameter tuning
4. Fixing Outliers in Residual Plot
5. More features like festivals, weather, etc

Thank You