A

PROJECT REPORT

ON

FOOD DEMAND FORECASTING FOR FOOD DELIVERY COMPANY

SUBMITTED TO

ORACLE ACADEMY

BY

CHINTAMADKA LAKSHMI

MAY 3rd,2021

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

JNTUH COLLEGE OF ENGINEERING SULTANPUR

PULKAL(M),SANGAREDDY(DIST

Repository:
https://github.com/chintamadkaLakshmi/food_demand_forecast for food delivery company

TRAINING GUIDE:

Ashok Kumar Sir

Hemant Kumar Sir

Pradepthi Mam

Sreeja Vasala Mam

PROJECT GUIDE:

Hemant Kumar Sir(SmartBridge)

P.Krupa Sagar Sir(Jntuhces)

CONTENTS

ABSTRACT

"Demand is an economic principle referring to a consumer's desire to purchase goods and services and willing pay a price for a specific good or service".

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 pershiable 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 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.

PROBLEM STATEMENT

Client : meal delivery cmpny

Problem:

-> Deals with a lot of pershiable materials

-> Not enough inventory->out-of-stock->push customers to competitors

-> Too much inventory->more wastage of food

Solution will also help in:

-> Planning the stock of raw materials

-> Staffing of the centers

Pre-Requisite document for Java:

Hardware Requirement:

Windows 7 and above (64-bit)
RAM: 4GB
Processor: Minimum Pentium 2 266 MHz processor
Browsers: Chrome

Software Required:

Java JDK 10
Weka
Eclipse IDE

STEPS IN DATA PREPROCESSING:

step1: Import the libraries
step2: Import the data-set
step3: Check out the missing values
step4: Splitting the data-set into Training and Test Set
step5: Feature Scaling

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	UniqueID for Fulfilment
meal_id	UniqueID for meal
checkout_price	Final pice including charges
base_price	Baseprice of the meal
emailer_for_promotion	E-Mailer sent for promotion
homepage_featured	Meal featured at home
num_orders	(Target)OrdersCount

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.

Feature Extraction

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.

Feature Extraction

VariableName	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

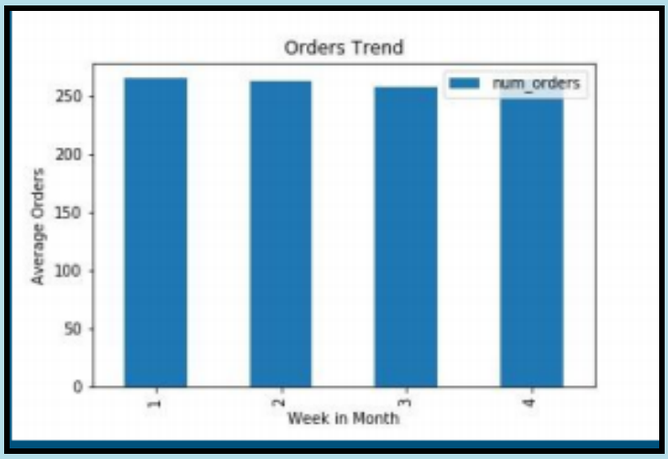
Deriving new variables by grouping consecutive weeks into one parent class.

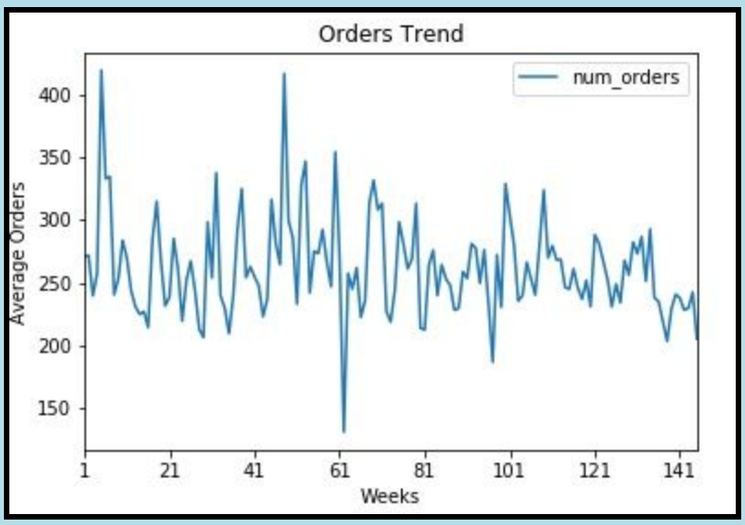
Variable	Description
year	It represents the year, group of 52 consecutive weeks, in which the record belongs.
month	It represents the month, group of 4 consecutive weeks in a year, in which the record belongs. Since, month is considered as a set of 4 weeks, there are 13 months in the dataset.
quarter	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 rec) ● 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

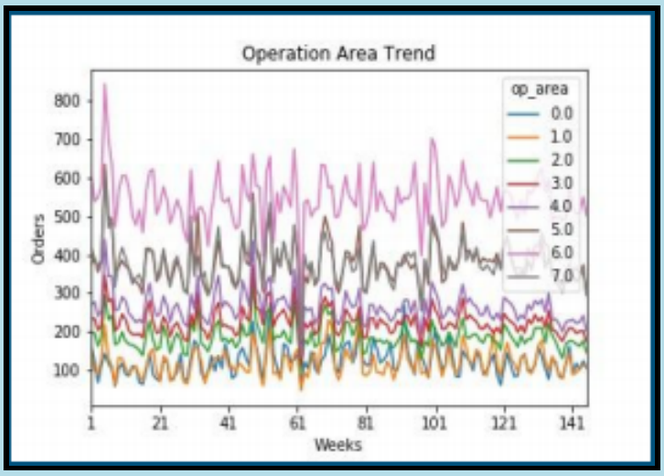
Overall Orders Trend

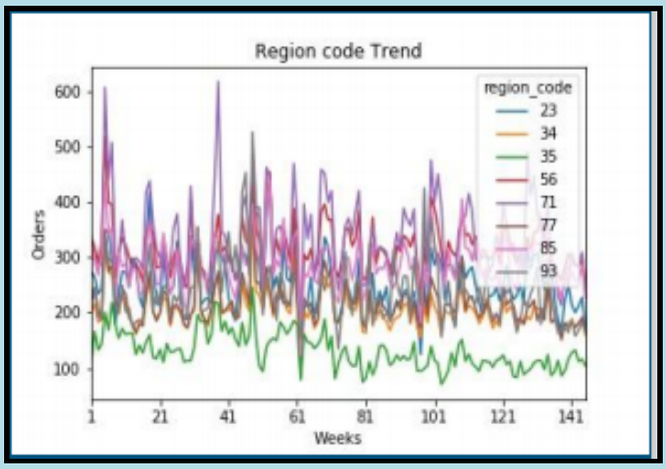


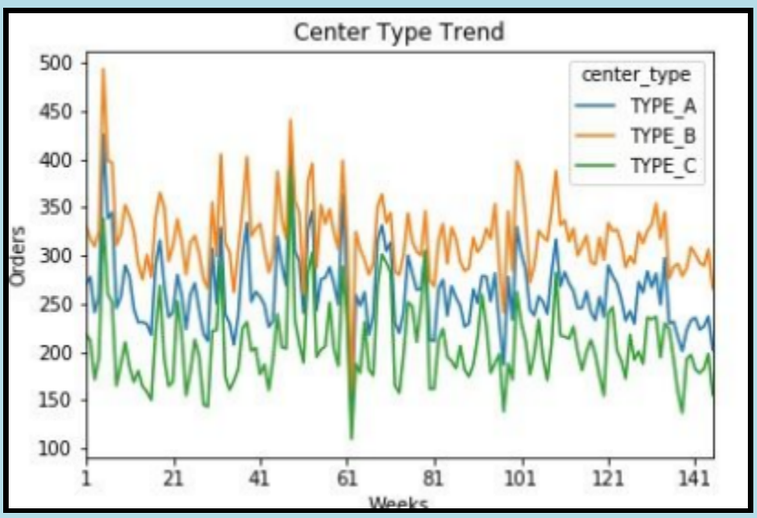


Above plots represents the Monthly, Week in month and weekly Orders Trend respectively. Below are findings from above plots:

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.



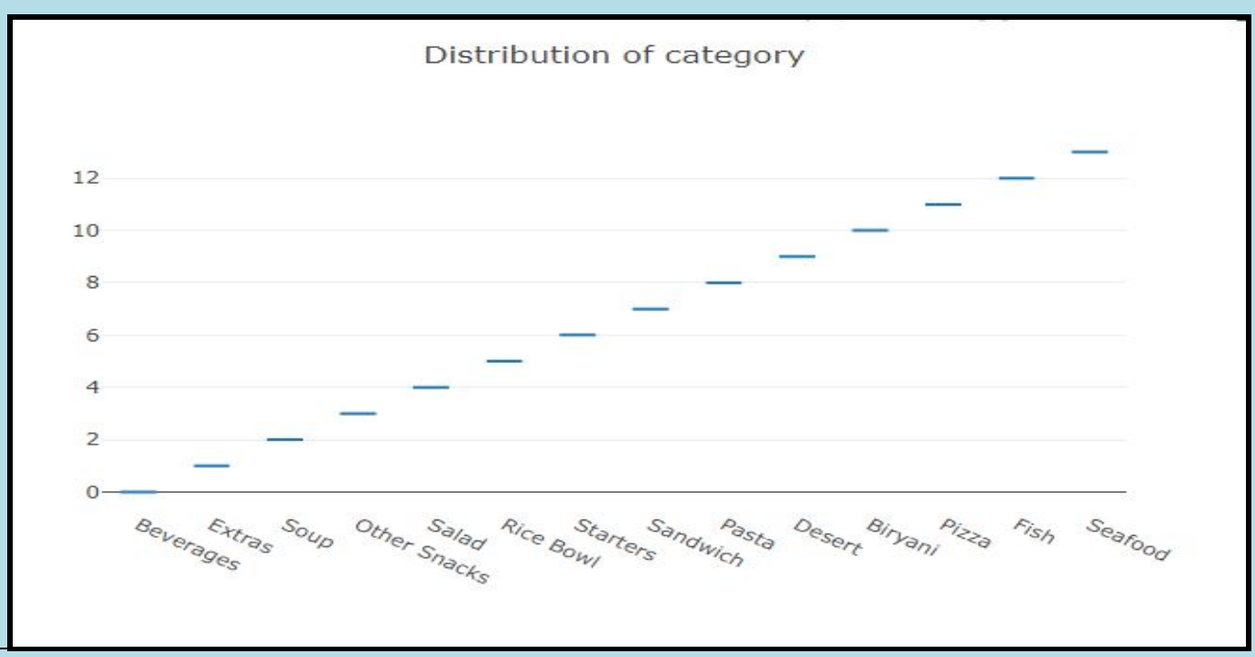


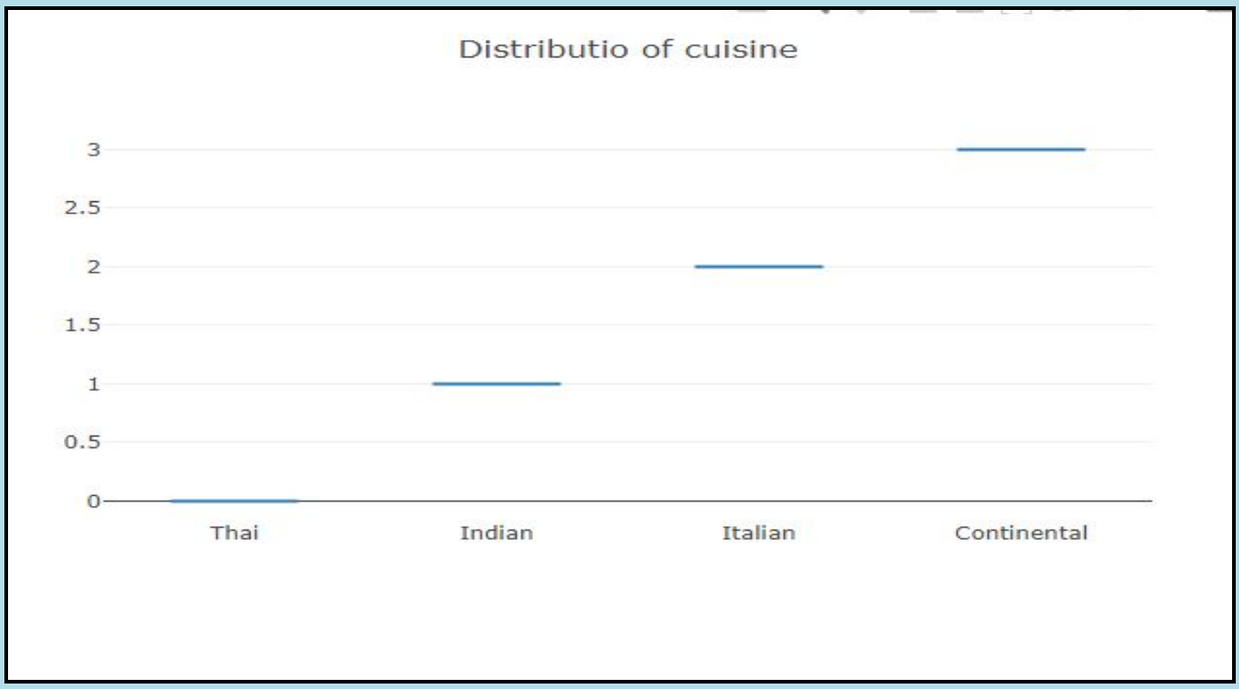


Above plots represents the weekly order trend with respect to the center's operation area, region code and center type respectively. Below are findings from above plots

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



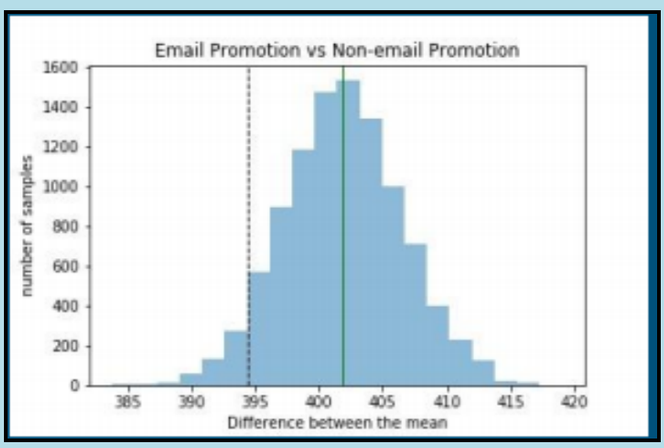


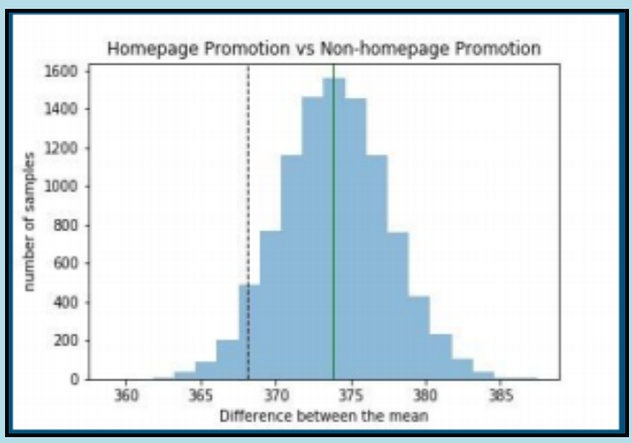
First plot represents the weekly orders trend in meal category and second plot represents the same in cuisine. Below are findings from both the plots:

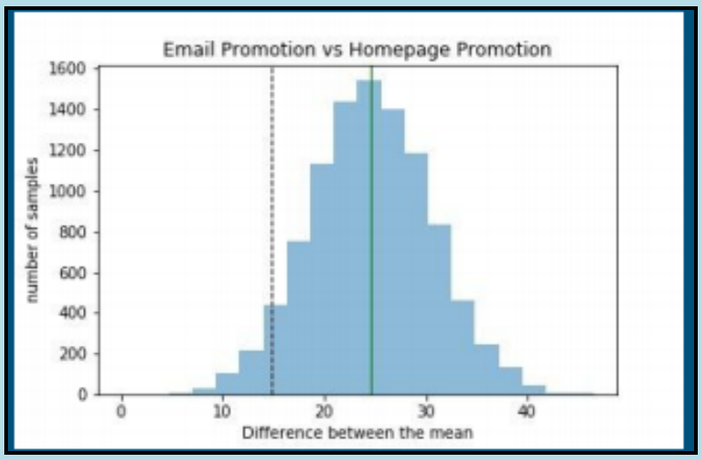
1. Orders for Italian meals and Beverages are always high
2. Orders for Salad increased after week 18
3. There are fluctuations in the number of orders for Indian meals, Rice Bowl and Sandwich

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?



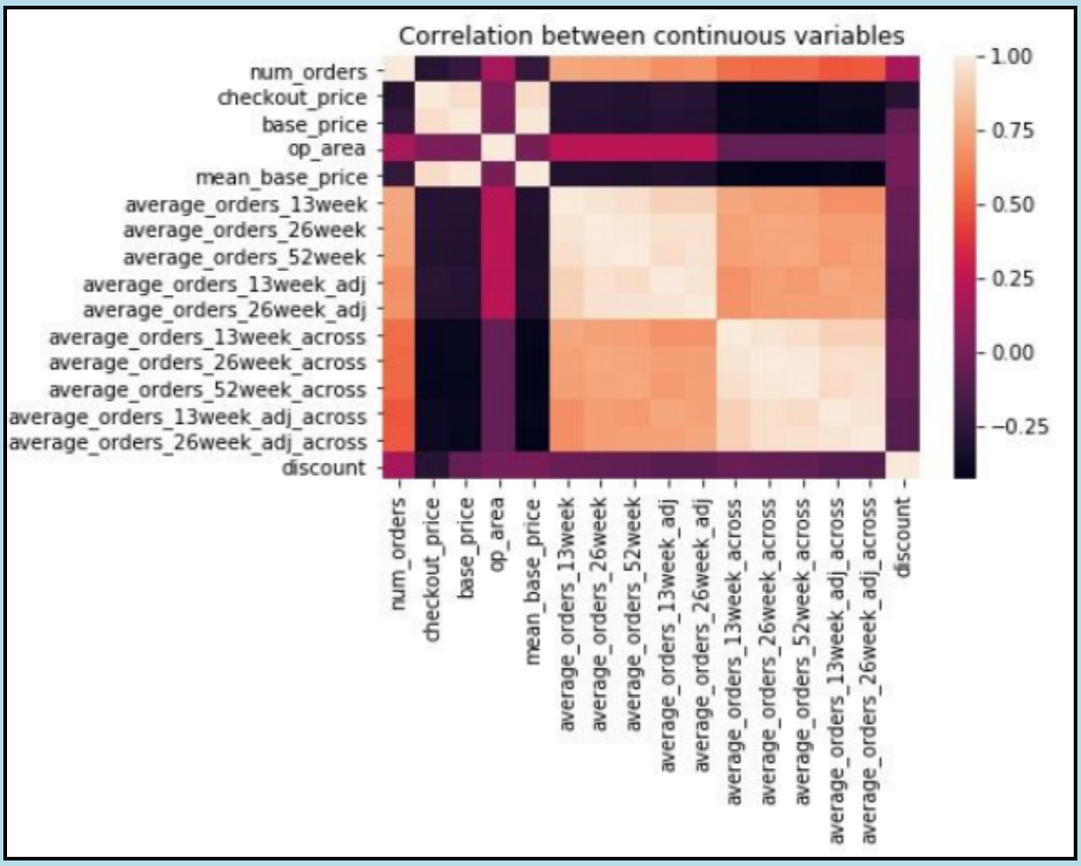




All the above questions were answered using the hypothesis test. The distribution of difference in mean of number or orders, one distribution for one question, are displayed above. Below are outcome of tests

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

Feature Engineering

The feature extraction was done during the data wrangling section. There are mainly 2 types of features

- **Continuous**: average_orders_Nweek, average_orders_Nweek_adj, average_orders_Nweek_across, average_orders_Nweek_adj_across, checkout_price, base_price, mean_base_price, discount, op_area
- **Categorical:** week, center_id, meal_id, emailer_for_promotion, homepage_featured, region_code, center_type, city_code, category, cuisine, year, month, quarter, week_in_month

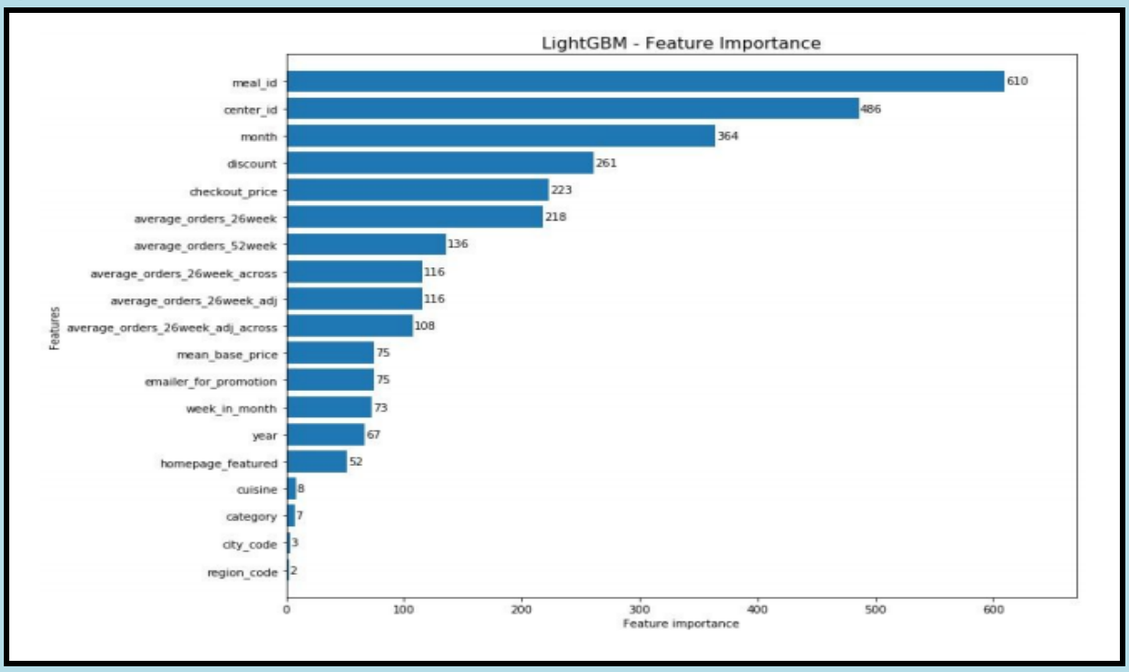
After exploratory data analysis, it was found that below features are not useful enough in prediction:

- **base_price:** Since, the change in base_price can also impact the orders, mean_base_price is the better representation of actual price than base_price.
- **quarter:** month can be considered as granular version of quarter and hence, model will do better with month.
- average_orders_13week, average_orders_13week_across: Since, testset will not have complete information on past 10 orders, it is not accurate feature.
- **week:** The training data contains weeks from 1 to 145 and test data contains week from 146 to 155. Hence, model will not work properly with week variable.

Further filtering of features was done using the out-of-box LightGBM model. Since, it is faster to train and can handle categorical features easily.

In the first Iteration, many models were trained using randomly selecting the features. On evaluating them on testset,it was found that model trained withoutaverage_orders_13week_adj, average_orders_52week_across, average_orders_13week_adj_across features performs better.

In the next Iteration, Model was trained using remaining features, and feature importance was analysed. Below plot displays the importance of each feature in the model. Note that, op_area have no importance in this model, hence also remove that.



Finally, center_id, meal_id, checkout_price, mean_base_price, discount, emailer_for_promotion, homepage_featured, city_code, center_type, category, year, region_code, month, week_in_month, cuisine, average_orders_26week_adj, average_orders_52week, average_orders_26week, average_orders_26week_adj_across, average_orders_26week_across features were selected for the next iteration.

Modelling

The Client, meal delivery company, wants to forecast the orders for upcoming weeks. This is a regression problem were model needs to predict the num_orders on input week for a product-center combination. For this problem statement, we will select lightGBM and XGBoost as the algorithms. Many experiments targeting specific perspective were performed.

Target Variable

Target variable, num_orders, which will be the outcome of the model, should be non-negative natural number. In order to restrict this in the model, natural logarithm of the num_orders at the time of training is passed to the model and exponential of the model's outcome is rounded to the nearest integer.

Outliers

One experiment was to check the performance of model with and without the outliers. It was found that model without outliers performed better on validation set but bad on test set.

Model	Validation Set	Test Set
Outliers	46.6043	51.0826
Without Outliers	46.3979	51.3646

Label Encoding vs One Hot Encoding

Dataset contains many categorical variables and some contains value in the form of String. Algorithm does not accepts the variable in the form of string and even does not understand the difference between categorical and continuous variables. Hence, after trying both methods, it was found that One Hot Encoding performed better than Label Encoding in the testset

Models	Validation Set	Test Set
Label Encoding	46.6043	51.0826
One Hot Encoding	47.3017	51.0484

Continuous Non-negative Variables

Dataset contains features that are non-negative continuous variables. Experiment was performed by passing values as it is and also by taking the natural logarithm of values. It was found that passing values directly performs better than natural logarithm.

Models	Validation Set	Test Set
Raw values	46.6043	51.0826
Natural Logarithm	47.3232	51.4003

LightGBM vs XGBoost

Model was trained using two algorithms, viz. LightGBM and XGBoost. After comparing the performance of both, it was found that LightGBM performs better.

Models	Validation Set	Test Set
LightGBM	47.3017	51.0484
XGBoost	50.2592	52.7788

Hyper-parameter Tuning

Until now, models were using default parameters. After, tuning the parameters of both the algorithm. It was found that both the models were performing nearly the same.

Models	Parameter	Validation Set	Test Set
LightGBM	num_leaves: 51 n_estimators: 260 min_child_samples: 45	44.6960	50.5356
XGBoost	missing:0.0 max_depth: 9 n_estimators: 300 min_child_weight: 45	43.4390	50.5686

Ensemble

Since, both the models are performing nearly the same on test set. Average of outcomes from both the models was also evaluated on test set. It gave the result of 50.2260.

Advantages :

- Food demand forecast is one of the major problem in today's world ,based on demand forecasting we can solve this issue.
- High service levels, require enough stock
- **wide variety:** The oline menu provides an option for the user to select various restataurants.The images of cuisine and dishes attract a user to order.
- From a recent report,75% of millennials are interested in utilizing offers from a food delivery service.

- A food delivery app offers a different mode of payment like a Credict card debit card ,net banking ,and COD.Awallet option also provided to help user pay for their order.

Disadvantages:

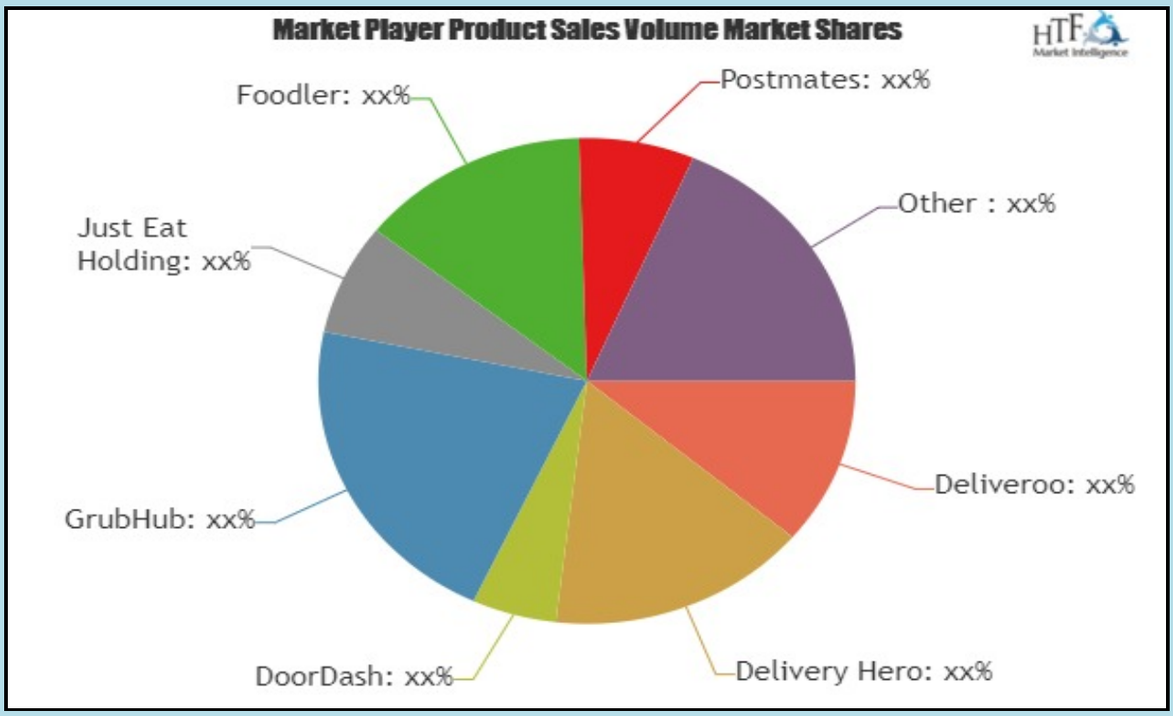
- when demand forecasting is not critical to the the business,simple models will suffice.But when the operational decisions made are based on it,having a robust and accurate model is key
- Too much stock

  Involves distresssed product wastage or discounted.
- Low stock

  Involves disastified customers missed sales and bad reputation

Applications

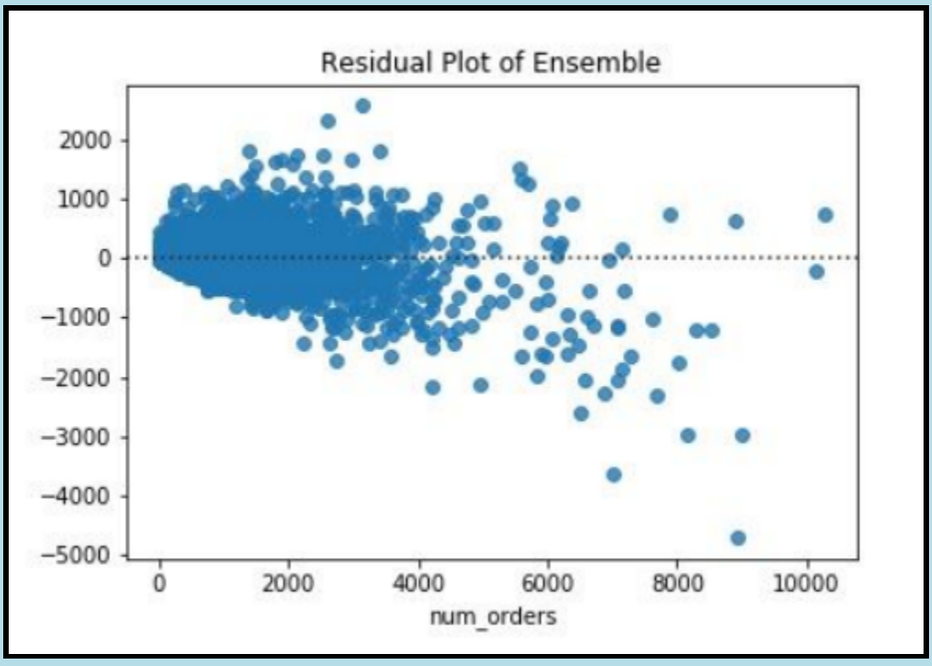
A few of the key players opration in the global online on demand food delivery services market are • Deliveroo • Delivery Hero SE • FoodPanda • Glova • Grubhub • Swiggy • Takesway.com • Zomato Mesdia Pvt.Ltd • Uber Technologies	



Conclusion

After performing different experiments with features, data format, algorithms, parameters, etc. The ensemble of tuned lightGBM and XGBoost performs better than others. The evaluation metric of 100 * RMSLE final model is getting the score of 50.2260.

Below is the residual plot of the final model.



Further Improvements can be done to improve the model. Below are some ideas that can be explored.

- The focus was on features related to meals. Features related to the centers can also be thought and try.
- Only LightGBM and XGBoost algorithm were used, other algorithms can be explored.
- Hyper-parameter tuning might lead to local minima, further tuning can be tried for better results.
- Residual Plot was not analysed. It can be analysed and fix the models for the issues.

APPROCHES TO IMPROVE THE PROJECT OUTCOME

- More features related to the outcomes
- Fixing Outliers in Residual Plot
- We can also try to bulid model with more different regression models.
- More features like festivals ,weather,etc

REFERENCES:

- Ballou R.H.2007.'The evolution and future of logistics and supply chain management'.European business review,Vol.10,pp.332-48.
- Alexandrov T..Bianconcini S..Daqum E.B..Maass P.and Mcelrov T.S 2012 'A review of some modrn approches to the problem of trend extraction'.Econometric Reviews,Vol.31,pp.593-624.