Hemachandar Nagarajan

Food Demand Forecasting

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

Demand Forecasting is a process by which an individual or entity predicts the how much the consumer or customer would be willing to buy the product or use the service. Without Proper Demand forecasting it becomes impossible for any business to function. Improper Demand forecasting. would result in heavy loss. Different industry or company has different methods to predict the demands. In case of food industry, it is at most important that the demand needs to be on bulls' eye since the food materials gets perished easily and has the fixed time frame to be used. So, the daily and weekly demand needs to be precise to avoid wastage which would otherwise increase the operating cost.

Problem Statement

The data set is related to a meal delivery company which operates in multiple cities. They have various fulfilment centers in these cities for dispatching meal orders to their customers.
The dataset consists of historical data of demand for a product-center combination for weeks 1 to 145.
With the given data and information, the task is to predict the demand for the next 10 weeks (Weeks: 146-155) for the center-meal combinations, so that these fulfilment centers stock the necessary raw materials accordingly.

Business Benefits

The replenishment of raw materials is done only on weekly basis and since the raw material is perishable, the procurement planning is of utmost importance.
Therefore predicting the Demand helps in reducing the wastage of raw materials which would result in the reduced cost of operation. Increased customer satisfaction by timely fulfilling their expectations and requirements.

Hemachandar Nagarajan

Data Dictionary

The dataset consists of three individual datasheets, the first dataset contains the historical demand data for all centers, the second dataset contains the information of each fulfillment center and the third dataset contains the meal information.

Weekly Demand data (train.csv):
Contains the historical demand data for all centers. The Train dataset consists of 9 variables and records of 423727 unique orders. test.csv contains all the following features except the target variable. The Test dataset consists of 8 variables and records of 32573 unique orders.

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	Emailer sent for promotion of meal
homepage_featured	Meal featured at homepage

Variable	Definition
num_orders	(Target) Orders Count

fulfilment_center_info.csv:
Contains information for each fulfilment center. The dataset consists of 5 variables and records of 77 unique fulfillment centers.

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)

meal_info.csv:
Contains information for each meal being served

Variable	Definition

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

Libraries Used

pandas,

numpy,

scikit learn,

matplotlib,

seaborn,

xgboost,

lightgbm,

catboost

Data Pre-Processing

- There are no Missing/Null Values in any of the three datasets.
- Before proceeding with the prediction process, all the three datasheets need to be merged into a single dataset. Before performing the merging operation, primary feature for combining the datasets needs to be validated.
- The number of Center IDs in train dataset is matching with the number of Center IDs in the Centers Dataset i.e 77 unique records. Hence, there won't be any missing values while merging the datasets together.

- The number of Meal IDs in train dataset is matching with the number of Meal IDs in the Meals Dataset i.e 51 unique records. Hence, there won't be any missing values while merging the datasets together.
- As checked earlier, there were no Null/Missing values even after merging the datasets.

Feature Engineering

Feature engineering is the process of using domain knowledge of the data to create features that improves the performance of the machine learning models.

With the given data, We have derived the below features to improve our model performance.

- Discount Amount : This defines the difference between the “base_Price” and “checkout_price”.
- Discount Percent : This defines the % discount offer to customer.
- Discount Y/N : This defines whether Discount is provided or not - 1 if there is Discount and 0 if there is no Discount.
- Compare Week Price : This defines the increase / decrease in price of a Meal for a particular center compared to the previous week.
- Compare Week Price Y/N : Price increased or decreased - 1 if the Price increased and 0 if the price decreased compared to the previous week.
- Quarter : Based on the given number of weeks, derived a new feature named as Quarter which defines the Quarter of the year.
- Year : Based on the given number of weeks, derived a new feature named as Year which defines the Year.

Data Transformation

- Logarithm transformation (or log transform) is one of the most commonly used mathematical transformations in feature engineering. It helps to handle skewed data and after transformation, the distribution becomes more approximate to normal.
- In our data, the target variable ‘num_orders’ is not normally distributed. Using this without applying any transformation techniques will downgrade the performance of our model.
- Therefore, we have applied Logarithm transformation on our Target feature ‘num_orders’ post which the data seems to be more approximate to normal distribution.
- After Log transformation, We have observed 0% of Outlier data being present within the Target Variable – num_orders using 3 IQR Method.

Evaluation Metric

Hemachandar Nagarajan

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.

Initial Approach

- Simple Linear Regression model without any feature engineering and data transformation which gave a RMSE : 194.402
- Without feature engineering and data transformation, the model did not perform well and could'nt give a good score.
- Post applying feature engineering and data transformation (log and log1p transformation), Linear Regression model gave a RMSLE score of 0.634.

Advanced Models

- With improvised feature engineering, built advanced models using Ensemble techniques and other Regressor algorithms.
- CatBoost and LightGBM Regressors performed well on the model which gave much reduced RMSLE.
- With proper hyper-parameter tuning, CatBoost Regressor performed well on the model and gave the lease RMSLE of 0.5237
