BEFD Project

Food Demand Forecasting

The business problem:

A meal delivery company operates in multiple cities. They have various fulfillment centers in these cities for dispatching meal orders to their customers. We need to provide these centers the demand forecasting for upcoming weeks so that these centers will plan the stock of raw materials accordingly.

Task:

Predict the demand for the next 10 weeks!

Data sources

Datasets

- Fulfilment center data
- Meal info data
- Sales historical data

Fulfilment centers data features

Variables:

- **center_id**: Fulfilment identifier
- **city_code**: City id in which the center is located on
- **region_code**: Region id in which the center is located on
- **center_type**: Type of the center
- **op_area**: Size of the operational area

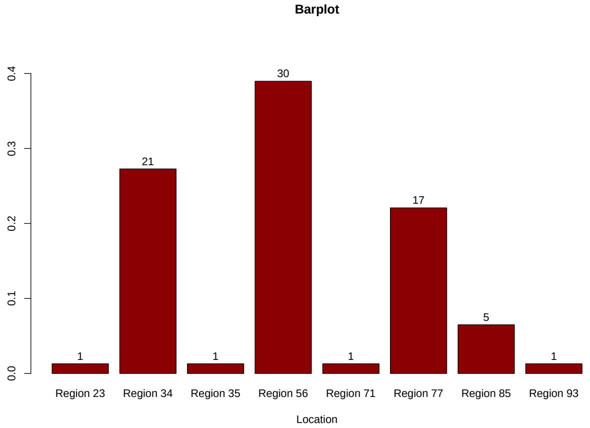
Fulfilment centers data features summary

Variables:

- **center_id**: We have 77 different centers
- **city_code**: The company acts on 51 different cities

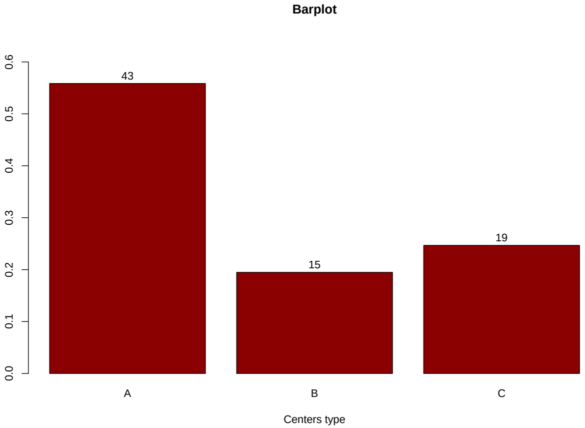
Fulfilment centers data features summary

region_code: These cities are located on 8 different regions



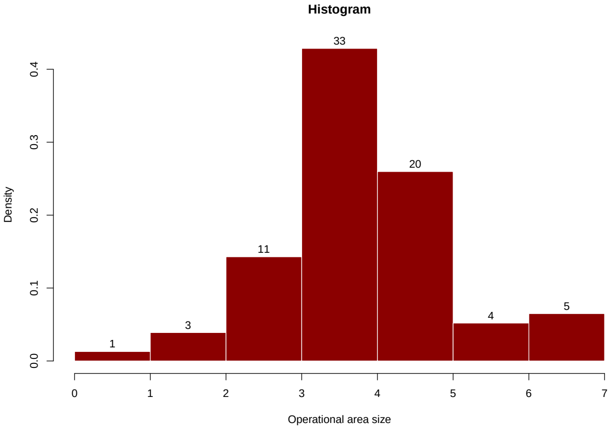
Fulfilment centers data features summary

center_type: There are 3 centers types: A, B and C



Fulfilment centers data features summary

- **op_area**:



Fulfilment centers data rows

center_id	city_code	region_code	center_type	op_area
11	679	56	TYPE_A	3.7
13	590	56	TYPE_B	6.7
124	590	56	TYPE_C	4.0
66	648	34	TYPE_A	4.1
94	632	34	TYPE_C	3.6
64	553	77	TYPE_A	4.4

Meal info

meal_id	category	cuisine
1885	Beverages	Thai
1993	Beverages	Thai
2539	Beverages	Thai
1248	Beverages	Indian
2631	Beverages	Indian
1311	Extras	Thai

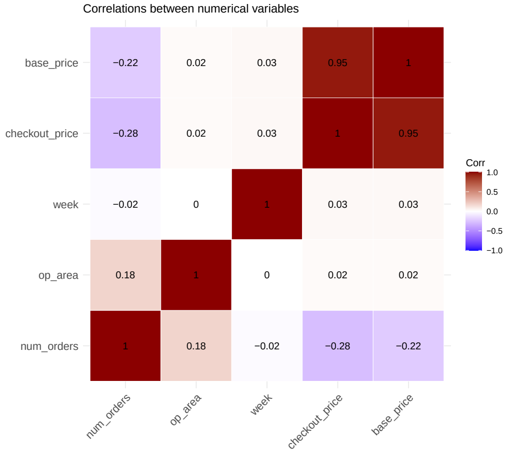
Sales data

id	week	center_id	meal_id	checkout_price
1379560	1	55	1885	136.83
1466964	1	55	1993	136.83
1346989	1	55	2539	134.86
1338232	1	55	2139	339.50
1448490	1	55	2631	243.50
1270037	1	55	1248	251.23

Sales data

id	base_price	num_orders
1379560	152.29	177
1466964	135.83	270
1346989	135.86	189
1338232	437.53	54
1448490	242.50	40
1270037	252.23	28

Correlations



Modelling

Since we want to organize the goods for each specific fulflment center, we need to forecast the demand for each specific center. Moreover, we also need to stratify for each unique meal, since each of them requires a different set of raw materials. We propose a two-stage approach:

- First we account for the temporal relationship using the linear model, obtaining (hopefully) i.i.d. residuals
- Then we model the obtained residuals, using some flexible method such as the gradient boosting

Linear model

We want to fit a straight line, between demand and time, for each combination of center and meal. This mean we should fit $N^o centers \cdot N^o meals$ $(77 \cdot 51 = 3927)$ linear models. But if we carefully craft some indicator variables we can specify all the simple linear models in to one single big linear model.

Linear model

$$Y_{ij} = \beta_{0ij} + \beta_{1ij} week + \mathcal{E}_{ij}$$

$$\forall i = 1, ..., 77; j = 1, ..., 51$$

Is equivalent to:

$$Y = \beta_0 + \beta_1 week + X_{ind}\beta_{level} + X_{ind}\beta_{slope} \cdot week + \mathcal{E}$$

Linear model

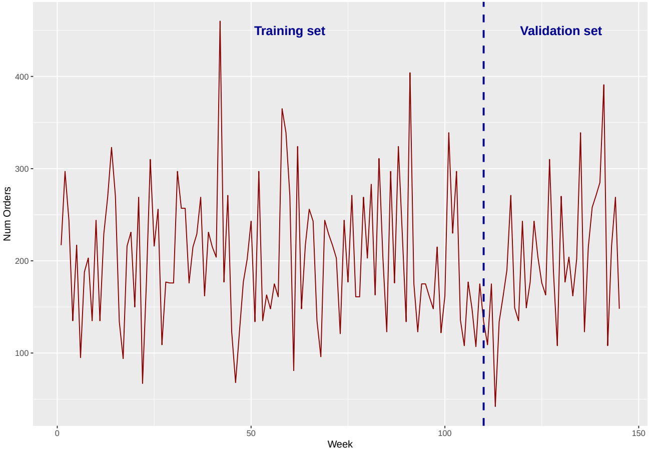
where X_{ind} is a vector with $51 \cdot 77 - 1 = 3926$ columns, and is obtained as the interaction between the dummy expansion of the categorical variables center_id and meal_id.

The model has $1 + 1 + 3926 + 3926 = 7854$ scalar parameters, that in the simple formulation there are 2 parameters for each model, so $2 \cdot 77 \cdot 51 = 7854$

Validation set and Test set

Dealing with time series data means that standard cross validation is not a viable option, since it breack the temporal dependency. We instead reserved a validation set, taking the last set of observations. The test set are the next 10 week, and the true number of orders stands on Kaggle.

Validation set and Test set

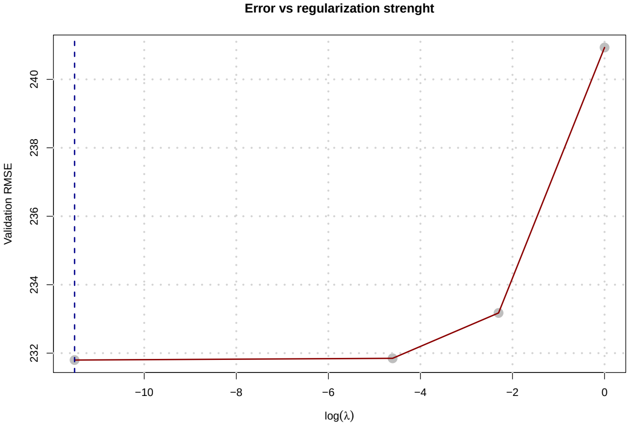


Regularization

We added elastic-net regularization in the estimation process:

$$\hat{\beta} = arg \min_{\beta \in \mathbb{R}^p} \left(\sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right) \right)$$

No regularization is needed



Results on validation set

Model	RMSE	MAE
Mean	350.9743	211.5107
LM	240.9311	106.4178
LM on ln(y)	367.1824	181.3026