**Sales Acquisition model (in progress)**

* **M.Kaleeswaran, Analytics Intern (Swiggy, NCR), under Mr. Vinit Kumar**

**Objective:**

To make closest predictions on restaurant orders and AOV over a time period (model trained for 2 months), to get a general perspective on GMV of each restaurant and acquire restaurants more efficiently going by this list.

**Basic outline:**

Conduct regression modelling separately, setting the **number of orders** and **AOV** as target variables and make predictions on the values based on our predictor variables: Cuisine, **CFT, Area, Zomato Votes \* Zomato Ratings.**

**Note:**

These are the parameters under consideration for basic modelling to understand how the data works. The more reasonable parameters we add to it, much better the predictions would be.

Parameters:

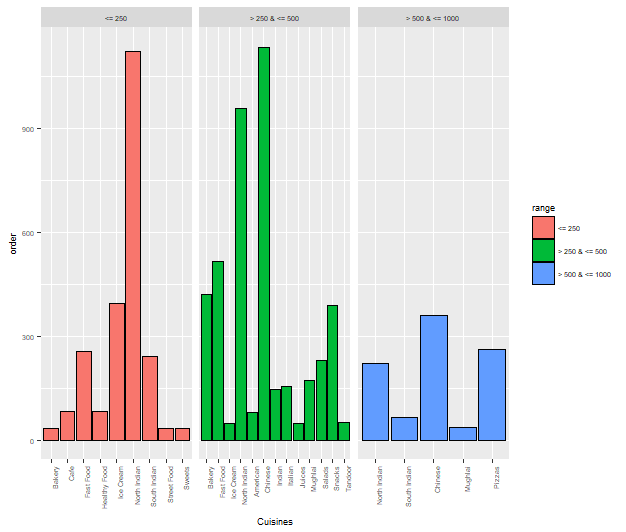
* Cuisines: To get relative dependency of orders and AOV on various Cuisines.
* CFT: To get relative dependency of orders and AOV on various CFT ranges.
* Area: Since each CFT ranges and Cuisines have different order generation on different areas, an interaction term of area with all other variables will be created.
* Zomato Votes \* Zomato Ratings: It might happen that if two restaurants have same CFT, Cuisine and belong to same area and still these two restaurants would generate different number of orders. Zomato Votes \* Zomato Ratings, might be one if the factors that might explain this variability.

**Execution and workflow:**

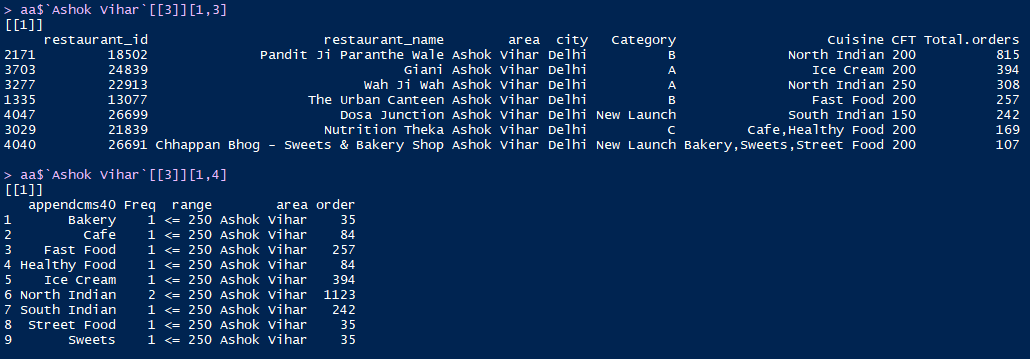
The whole model was developed using R programming. Initially to have an understanding of the demand on various Cuisines and CFT ranges in different areas, a demand analysis has been done on top 40 percentile orders generating restaurants, since this constituted for over 90 percent of the total orders. The following are the summary of the compiled data on Ashok Vihar:



Shows the CFT range, number of restaurant – a, number of orders – b.



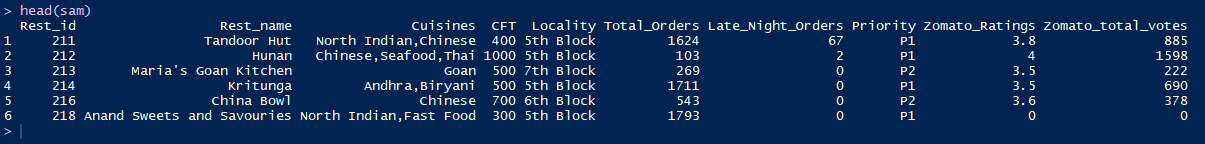
Shows the distribution of orders with different cuisine under different CFT ranges in Ashok Vihar



Shows the restaurant belonging to <= 250 CFT range and its details and also the number of orders coming from each of the cuisine in the CFT range. The same details has been summarized on all the CFT ranges and all the zones.

Now coming to the workflow process, we will train the model on the data containing the Total orders, Cuisines, CFT and Area for all the restaurants Swiggy Restaurants in India and make predictions on the restaurants listed in QOS sheet.

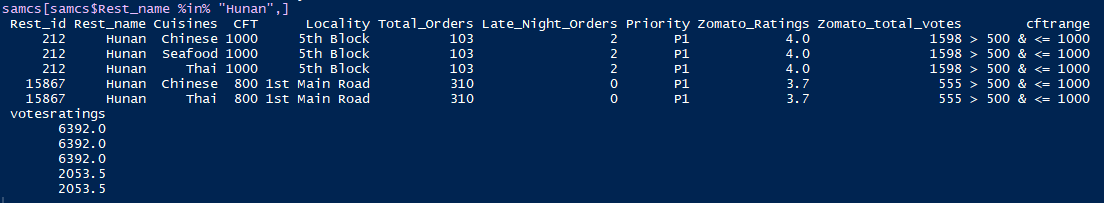
In order the make the data workable a lot of transformations has been made to the data, which will be detailed here:



First six entries of our data. Gives an idea about the variables under consideration.

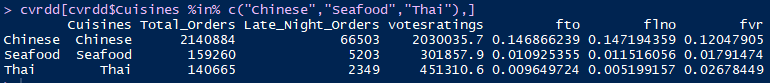
I’ll brief the data processing with an example, as the Hunan (Rest\_id = 212).

Since we are going to consider each cuisine separately, we should split a single restaurant with multiple restaurant as multiple restaurant of single cuisine.

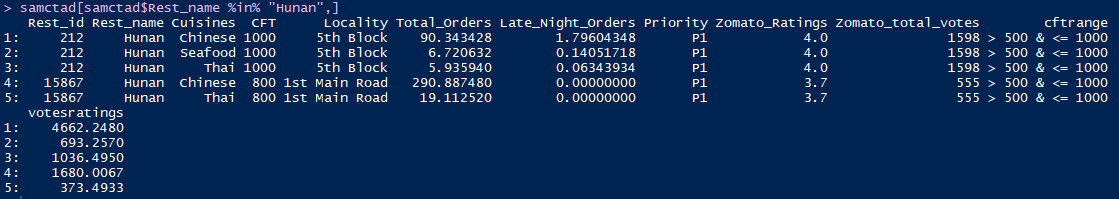


Hunan single restaurant of multiple cuisines has been split into multiple restaurants of single cuisine

Now since we have split a single restaurant in many, it doesn’t make sense to say each of had done the same number of orders. So we create a distribution on the fraction of orders generated from each cuisine and distribute it into each restaurant accordingly, similarly we do the same on the votes \* ratings, since a restaurant famous for one cuisine needn’t be on another.



Our distribution on the Cuisines of our sample case “Hunan”



New values of orders and votes \* ratings based on the distribution

Calculations based on the distribution for Hunan restaurant.

Similarly the calculations can be made on other cuisines of Hunan, referencing from the distribution.

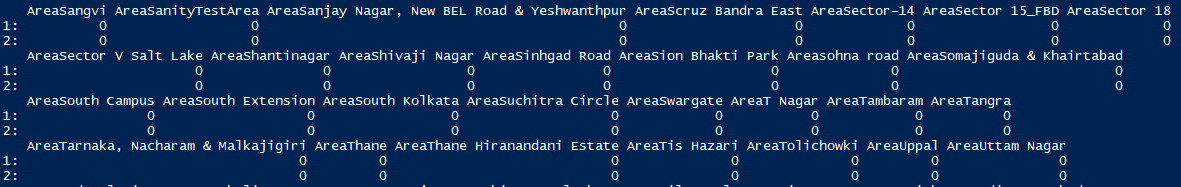
An initial iteration was made on the modelling taking the absolute values of the votes \* ratings and since it was producing a high variance between the estimates and the actual values, the votes \* ratings has been made into brackets. The logic of bracketing is as follows:

|  |  |
| --- | --- |
| Percentile of votes \* ratings | Bracket |
| 0 – 40 | 1 |
| 40 – 65 | 2 |
| 65 – 80 | 3 |
| 80 – 87.5 | 4 |
| 87.5 – 92.5 | 5 |
| 92.5 – 97.5 | 6 |
| 97.5 – 100 | 7 |

Now a regression model is been created based upon this data. After various iterations on the regression modelling, the best predictions as of now is been achieved with interaction terms of area with all of the other variables – CFT range, Cuisines and Votes \* Ratings bracket, since it would reflect an interplay between the area with its demand of the variables.

So the formula would look something like:

Now each of the predictor variables will be made into Dummy variables of its factors, taking integer values of 1 and 0 and the formula will be based on this.



Since whole of the data frame is too large to be printed, printing just a part of it.

Similarly we predict the AOV with the related data, only difference as of now would be a minor addition to formula in the case of AOV

This way GMV can be estimated:

**Current issues and later iterations to be worked upon:**

Though the model gives a good idea of how different restaurants fall in a scale of orders, it couldn’t get a fair estimate on very high order generating restaurants (the outliers). Need for more number of meaningful parameters. As more of reasonable parameters are been added the efficiency of the model could be increased. The data need to be trained on better algorithms, rather than simple linear regressions.