Project Submission | Fall 22

- Ashish Dhiman | ashish.dhiman@gatech.edu
- Abhinav Arun | aarun60@gatech.edu
- Anshit Verma | averma373@gatech.edu

Table of Contents

Summary:

Online groceries industry is valued at upwards of 3 Billion USD in India. With evolving customer tastes after the pandemic and rapid urbanization, there is a growing demand for quick and instant delivery of groceries. **Zepto** is one such player operating in this space in India and promises delivery of groceries within 10 mins. As part of this course project we want to look at the different Analytics and Data Science use-cases which can be leveraged to fulfil the 10-minute delivery promise of Zepto.

Background on Zepto :

Zepto is a start-up from Mumbai that promises to deliver groceries to people in under 10 minutes. The secret of the Zepto app is the network of cloud shops or micro-warehouses, and these help Zepto deliver orders within 10 minutes. Besides employing dark storefronts, Zepto also uses a one-of-a-kind product, Locus. This fantastic solution tracks clients' geostatistical data, traffic dynamics, and how much time will be required for the last mile delivery. In turn, the received data helps Zepto decide whether to build a new dark store in an area or not.

Zepto Website

Problem Definition:

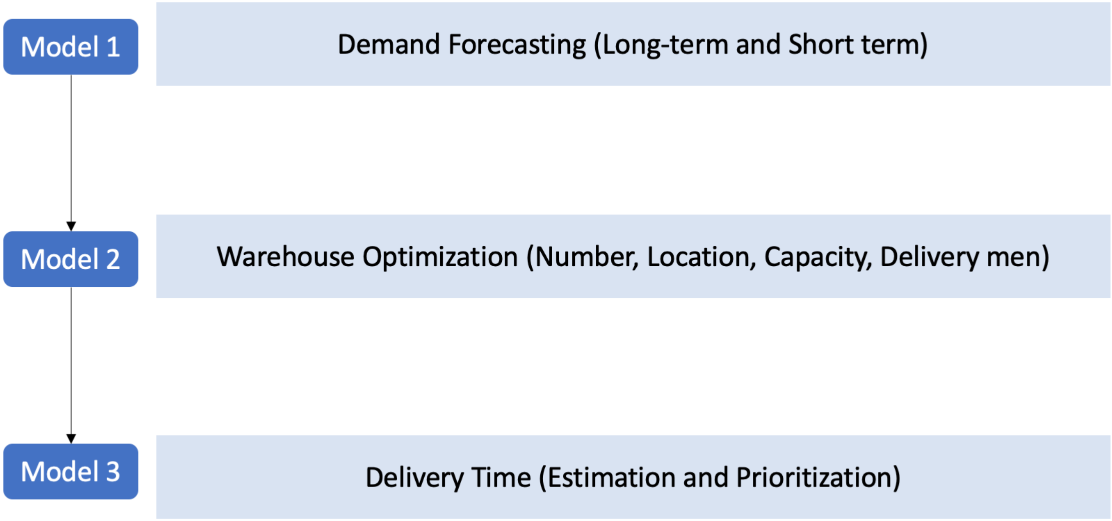
The problem is to work out a solution to a hyperlocal delivery system where we have to deliver majority of the orders within 10 minutes. To adhere to such stringent timelines and to build a USP, it is imperative that we leverage data driven technologies to work out a solution. The problem could be broken down into various hierarchical subproblems which we will look to address to come up with a sustainable and holistic solution.

These are the questions that we need to address before diving into data and models:

- How to forecast the demand of various grocery items? What all factors drive these? To simplify the problem, for now, we are assuming that we do not have any supply side constraints or any unusual macroeconomic activity affecting our planning.
  Addressing Multiple aspects of Demand forecasting:
  - Long Term Demand Forecasting: It would be used to drive growth tied objectives and to decide on phases to go out for expansion (cash burn cycle)
  - Short Term Demand Forecasting: Day to Day forecasting to help in operationalizing daily activities , working out optimal revenue generation streams and enhancing user satisfaction.
- How to use the above forecast to decide on locating the regions for expanding/setting up stores (dark front warehouses) that would help us adhere to our committed timeline of 10 minute delivery time. How to build a sustainable network of warehouses.
- At the same time, we need to estimate the capacity of warehouses, both in terms of volume and variety as we need to have a near perfect inventory management system in place to meet our objectives within a stipulated budget.
- After deciding on the same, we need to optimize the number of delivery guys to hire in order to meetup our forecasted demand with a budget constraint in place.
- The next question is to decide on the priority of the orders to be delivered. A lot of factors might be used in deciding on which order is to be prioritized. During instances of non-adherence, whom do we prioritize, users who are already onboarded and are regular user of our service or those users who are new to our platform?
- How to evaluate the data models build to address the above specified questions.

Analytics Methodology:

Some of the questions above can be answered independently while others are interdependent. To simplify the analytics problem, we have come up with the following **framework/pipeline**:



Model 1

Demand Forecasting (Long-term and Short term)

Model 2

Warehouse Optimization (Number, Location, Capacity, Delivery men)

Model 3

Delivery Time (Estimation and Prioritization)

Model Layer 1: Demand Forecasting

The first and the key layer in our analytics pipeline is a model to forecast demand of our products. We want to predict the demand of our SKUs at level of a block or a neighbourhood. We want to predict demand both at a granular level (say hourly or daily) as well as at a macro level (say a quarter or half yearly).

Hence **Given data** on:

- Economic status in a neighbourhood (can be proxied from the high end shops in the region etc, or utilizing google maps to track user activities)
- Demographic distribution in the neighbourhood, as younger people might be more inclined for online groceries (from Census etc)
- Internet Penetration in the area (from survey or telecom companies)
- Zoning of the area: Commercial or Residential (from Municipal)
- Distribution of homes in the area: single family homes or high rises etc,
- Population total and density etc. (from Census)
- Presence of other grocery stores in the neighbourhood (Google Maps API)
- Grocery Delivery apps being used by people (collected through survey)
- Seasonality and other patterns affecting demand
- Macroeconomic effects etc

We will **use models:**

- A combination of Time Series and Regression:
  Time Series part of the problem comes from temporal effects (trends, seasonality etc.) which will play into the Demand forecasting, while Regression will capture other non-temporal predictive effects.

We can use various Machine Learning algorithms like Linear Regression, Random Forest and XGBoost to solve the regression problem, we could use Shapley values to get the feature importance and narrow down the subset of features to features most predictive of the demand to take actionable business decisions. This way, we could inject interpretability to our model.

Concretely , we want **to predict** the following questions as part of this model:

- What will be the demand of SKUs in the different neighbourhoods (both short term and long term).

Model Layer 2: Warehouse Optimisation:

The second layer of the pipeline after getting the data for demand is to determine how the demand will be fulfilled. In order to determine that, we need to answer the following **questions**:-

1. How many warehouses should be built?
2. Where should the warehouse be located?
3. How many customers will a single warehouse serve?
4. What should be the area of the region that will be served by a single warehouse?
5. What should be the capacity of the warehouse?
6. What items should be stocked in the warehouse to ensure fulfilment of the order?
7. How many workers to hire in order to minimize the time it requires to fulfil an order?

In order to answer these question, we will use the ***data for demand*** we gathered in the previous step and combine it with various other ***data*** such as :-

- Geospatial Location of customers
- Average Traffic Data
- Weather Data
- Availability of Workers
- Time it takes to fulfil (pack) one order
- Items ordered by customers with their quantity

We can use the data and a combination of various ***models*** to answer the questions asked above.

- We can use linear regression model to answer question (1).
- We can use a combined optimization model to answer questions (2), (3), (4), and (5).
- We can use a linear regression model to answer question (6).

Model Layer 3: Delivery Time Prediction and Optimisation

The last modelling layer in our pipeline is the one that directly connects to the objective we defined at the start i.e., how to deliver online groceries within 10 minutes of the ordering time. With this modelling layer, we want to use inputs from preceding layers like demand forecast, warehouse location, and delivery resources to first estimate delivery time, and then use an optimization model to prioritize and schedule deliveries, such that the delivery time remains under 10 mins.

We also plan to involve the feedback from the last layer into the preceding layers such that we are able to better rationalise our solution. This will also help us to be proactive in opening new warehouses, in case the current resources are not able to fulfil the orders.

We want to answer questions such as:-

- How to prioritize the deliveries?
- How to schedule workers for delivery?
- How many delivery workers to hire?

Hence **Given data** on:

- Delivery Routes
- Traffic Data
- Geospatial Location of Customer
- Time it takes to make one delivery

We will **use models:**

- **An optimization model:**
  We will use an optimization to prioritize the deliveries such that it minimizes the delivery time. We can use distance and warehouse locations as a constraint.

- **A combination of optimization and linear regression model:**
  Such a combination will help to answer how to schedule the workers for delivery by optimization such that it minimizes the number of workers and if there is a requirements to hire more workers, we can use the linear regression model to predict the number of delivery workers we should hire.

Moreover, in terms of prioritization, we will use a analytical/optimisation framework, where a customer relatively far away is higher prioritised vs say a nearby located customer.

A tricky case to handle is to decide the prioritization of orders for a new customer who is onto the platform for the first-time vs a customer who is a regular user of the platform . Say , for simplicity we assume that the both the orders were identical, and it is only one order that we can cater at the moment? Then the question is whom should we prioritize?

The answer to this question is not straightforward and should be tested out through A/B testing over a period of time to gauge the effect because ideally it makes sense to prioritize the new onboarded user but sometimes the cost of a trusted customer churn is much more than the utility added by onboarding a new customer.

Thus, we can perform an A/B testing by randomly delaying or cancelling the order of both new and old customers at random and see the attrition / negative rate corresponding to each and then we can take a weighted decision in such scenarios going forward.

Conclusion:

The 10 minute grocery delivery problem is a complex one, and involves a lot of interdependent steps. Hence to simply the analytics framework, we have come up with a framework composed of three layers addressing individual parts involved. We have tried to balance granular details and simplicity, would love to receive your feedback on the above framework.