Leveraging Analytics for Supply Chain Optimization in Freight Industry

Kashyap Barua, Parikshit Barua, and Sandeep Agarwal

Abstract We live in a country whose logistics industry is slated to be worth $160 billion. Several start-ups have emerged in India trying to crack this globally yet unsolved problem statement. Moreover, the freight scene, being age old, has its segment of bottlenecks to deal with which are inclusive but not limited to problems like fragmentation, inflated costs, lack of visibility into lanes, limited digital capabilities and back hauling. Tech-driven start-ups provide solutions to overcome these challenges for the fleet owners in terms of optimized demand–supply matching, brokerage eliminations by connecting supply with relevant demand thereby reducing costs, expose truckers and fleet owners to unchartered lanes and also facilitate reverse loads to optimize costs for these businesses. This is where companies are leveraging data science and analytics to tackle these issues and help the businesses grow. Companies like Uber Freight, BlackBuck and Rivigo are using the best of technologies to monetize this industry. Data when logged in the right manner can help industries understand the intricacies of issues and help them overcome the same. A typical example of implementation would be using a simple regression technique to predict demand in a specific region so that supply can be exposed well within time in order to avoid idling period by these truckers. Tracking key metrics like supply turn around time (TAT), truck in transit duration (TiT), placement index and others can help organizations determine and optimize on these metrics to maximize revenue. Being convoluted of a system, this industry has been a tough nut to crack, especially in a country like India. This chapter discusses how companies set up the entire data platform and infrastructure, thereby facilitating the usage of data for advanced analytics techniques to solve some crucial supply chain problems in the freight industry. The

K. Barua (✉)
MiQ Digital India, Bengaluru, India
e-mail: kashyapbarua@gmail.com

P. Barua
KIIT University, Bhubaneswar, India
e-mail: parikshitb04@gmail.com

S. Agarwal
Axtria Inc, Bengaluru, India
e-mail: sandygarg65@gmail.com

S. Rautaray et al. (eds.), *Trends of Data Science and Applications*,
Studies in Computational Intelligence 954,
https://doi.org/10.1007/978-981-33-6815-6_11

chapter also talks about some of the use cases for analytics and machine learning to solve problems related to the freight industry. Firstly, we demonstrate some visualizations and representations as to how insights are drawn through analytics to solve these kinds of problems. We follow this up by discussing how data infrastructures are set up in organizations to collect freight data and then finally we showcase some ML techniques that are used in the freight sector of businesses. This in turn would help users understand the nuances of decision science and analytics with its capabilities in scaling businesses.

Keywords Freight industry · Analytics · Data science · Machine learning · Metrics · Supply chain

1 Introduction

Before we had a solid infrastructure to store and maintain data, the freight industry would be dependent on manual book entry in terms of data collection. Imagine keeping a record on a book about when a truck got booked, the truck arrived at a loading point, left the loading point and even fulfilled an order. This system of tracking is cumbersome, especially when you want to leverage analytics and data science techniques to upscale your business. With the advent of big data technologies and advanced data engineering techniques, the process of data collection and data audit has become even more simpler. Added to the above metrics/details, organizations have been collecting numerous data points to enhance their business output. They collect real-time GPS data about the trucks whereabouts, driver data through mobile applications suggesting each stop and at what location, user activity on the Load Board application and numerous other details. These organizations in turn provide shippers and carriers of every size, be it corporates or small and medium enterprises (SMEs), with actionable insights to build their supply chains even more efficiently than ever.

As shown in Fig. 1 [1], companies leverage data to track performance for their freight pickup and delivery performances. Here, it is observed that the overall freight ratings are the least in the northern region of India while on the contrary, the performances seem to be doing well in the Western region. The average ratings for pickups are 4.36 out of 5, and the same is 4.31 for the deliveries in the Western region encompassing states like Gujarat, Rajasthan, etc. Given that a vast majority of the trucking community come from the Western region of the country, it is evident that the same metrics would be well off for the same region. It is critical to track performance metrics across regions for any business case, to ensure smooth supply chain management.

From the analysis shown in Fig. 2, it is observed that FMC durables constitute almost 20% of all the products that have been moved in the month of June 2020. Other products that follow are paint, polyester, consumer durables, chemicals and oil. This is another instance wherein companies leverage data to keep track of

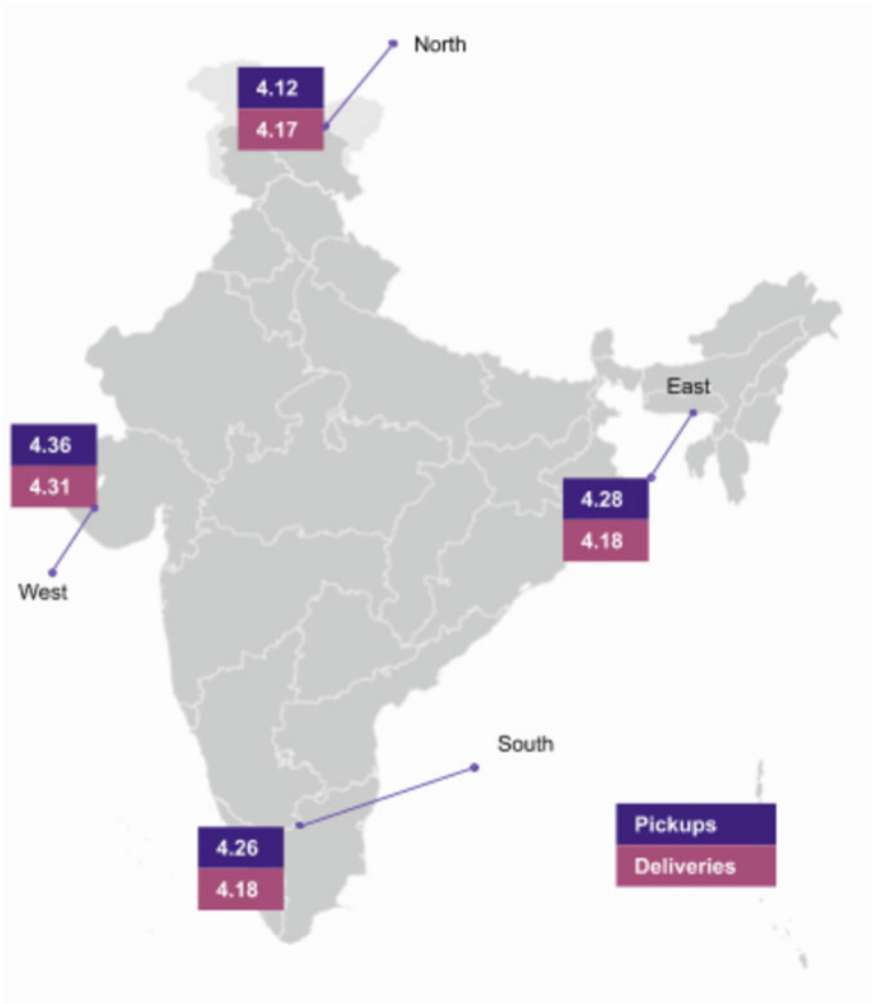

Fig. 1 Indian trucking report for pickup and delivery performance across regions

these nuances in their supply chain ecosystem. Demand and movement are about the type of product because the same type of trucks, be it half body, full body or standard containers, can be dispersed for the requirement with the clients. With the ever-increasing demands to ship products, the need to enhance this ecosystem will always increase. Hence, it becomes very crucial for the instrumentation of proper data warehousing infrastructures in organizations to draw insights and leverage data for optimized workflow.

Leveraging machine learning algorithms can help define a successful supply chain. For freight forwarders, implementation of these algorithms can be on a varied range of business fronts like prices, routes taken, volumes moved, customer requirements,

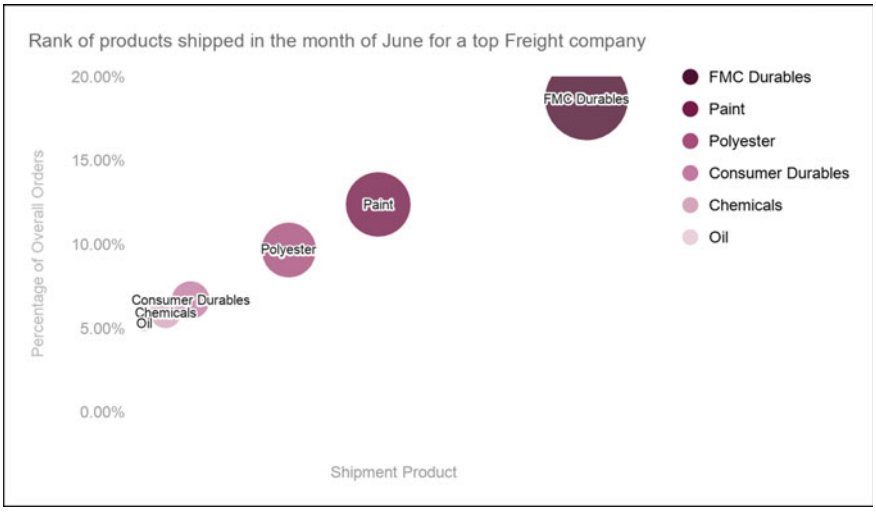

Fig. 2 Rank of products shipped in the month of June 2020 by a top freight company

etc. The data on these metrics can help optimize supply chains, yield business critical insights and processify operations. One of the key techniques used from machine learning is the technique of demand forecasting, which helps the demand side of the supply chain in anticipating demand across regions and businesses.

2 Literature Survey

In February 2020, Lóránt A.Tavasszy [2] emphasizes that the freight movement structure in the logistics industry is changing rapidly and the need to adapt to these developments has become even more important. He focuses on three basic scopes of the model improvement: the structural elements of the system that are modelled, the functional relations between these elements and the dynamic properties of these models. As per Lóránt, these innovation directions are independent of each other, and they help in reinforcing each other at various levels of the supply chain. At the end, he concludes that these areas need to be researched upon, to facilitate innovations at all levels which would facilitate in driving freight modelling research in the forward direction.

In October 2014, Paolo Ferrari [3], in his paper "The dynamics of modal split for freight transport", explained the dynamics of modal split in the transport system and the multimodality of the same. Also, the evolution of transport demand over the due course of time is also complimented by the evolutionary attributes of the modes of transport and the reaction of users to delay in cost variations.

In his chapter, Gregory Harris [4] focuses on improving the transportation and freight modelling by the implementation of a freight planning framework which leverages federal freight data and a variety of other tools and defines a model for statewide destination origin freight patterns, traffic models specific to freight movement and also system performance measures. His proposed FPF methodology considers data on freight flow or movements and structures the same into usable format for freight planning purposes at different levels. These types of model development and researches could be beneficial for the future supply chain industry.

Cruijssen Frans [5] discusses in his paper about the horizontal cooperation in the logistics and transportation sector. This method of horizontal cooperation has been an interesting topic of discussion from a theoretical standpoint of things, the reason being that it can be approached by personnel's coming from multiple disciplines like economics, operations and research. There is this collaboration between multiple businesses from different domains to ship products. A company like Flipkart or Amazon, who has their logistics department established can easily provide their services to SMEs and other businesses. This type of collaboration is quite evident in the country of India wherein these logistics dominators provide supply chain services to various other small-time businesses. This results in an ecosystem wherein each of these companies can benefit from the other in numerous ways. One of the reasons for the sustainability of this business model is because small businesses cannot start their own logistics services to transport goods, as oftentimes the system becomes complicated and expensive for them. This leads to the SMEs relying on the bigger players to help them ship their goods with sustained partnership and minimal costs, which benefits both the parties in their own ways.

3 Data Storage and Big Data Ecosystem

Leveraging analytics and data science techniques come with the responsibility of methodologies that an organization should adopt to collect and store data. This is an ever-increasing need, and companies have been on-boarding data engineers and data platform teams to upscale this aspect of technology. A data lake [6] is a centralized repository that allows one to store all their structured and unstructured data, at any scale. Any organization that is smart and proactively focuses on setting up these infrastructures will have a competitive advantage with their peers. These organizations with proper data lake implementations could do better predictions, customer segmentations and behaviour, understanding customer journey through clickstream in a much more efficient manner, thereby providing the competitive advantage in the business.

Companies have implemented data lakes wherein data from all sources and of all sizes are gathered in a single place. This facilitates processes such as reporting, visualizations, advanced analytics and machine learning to solve most business cases. A perfect example is a well-known tool called Qubole which is an open and secure data lake platform which enables machine learning, data streaming from different

sources and advanced analytics. The advantages of implementing a data lake as a data single point of data storage unit can be numerous, ranging like

- Collect and obtain quality data
- Support for all format of data
- Schema flexibility
- Democratize data
- Advanced analytics and machine learning
- Scalability (Fig. 3).

To highlight some of the key features depicted in the diagram above are as follows:

- Machine Learning—Data lakes allow organizations to leverage historical data to draw extensive insights and perform forecasting techniques to determine likely outcomes, which depends on historical data. This in turn provides a range of prescriptive solutions to achieve the optimal outcomes
- Advanced Analytics—Data lake architecture allows various professionals like data scientists, data analysts, business analysts to collaborate and access the

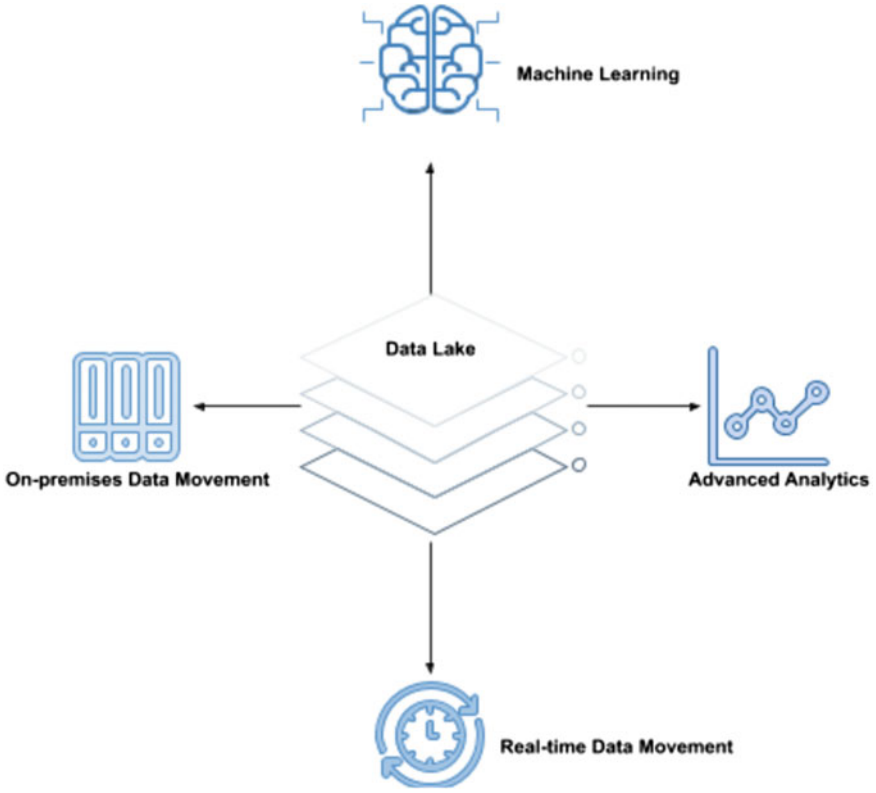

Fig. 3 Data lake key features

data in their choice of tools and framework to tackle a problem. Some of which includes frameworks such as Apache Hadoop, Presto, Apache Spark and other data warehousing tools.

- Real-time Data Movement—Data lakes facilitate the ability to stream/collect data in real time. This allows organizations in terms of saving time, scaling data to any size and saves the cumbersome task of defining schema, data structures and metadata information.
- On-Premises Data Movement—This allows to understand what data are present in the lake through methods of indexing, cataloguing and crawling.

4 Data Processing and Manipulation

One of the critical steps in a data science life cycle for any project would involve data processing [7] and manipulations (right after data collection and storage). This step involves shaping and cleaning the data as per business requirement or use case to solve problems.

The current trend for data capture and storage involves setting up data processing pipelines which pre-emptively cleans and shapes data even before ingesting them into storages. This enables analysts and data scientists to fast-track the analytics process as data cleaning/processing is a tedious task to perform (Fig. 4).

As per the diagram above, raw data can flow from multiple sources, be it real-time GPS data, survey data or salesforce data which is then pre-processed and streamed into the operational data storages. This is where the data quality assessment is done within the entire pipeline, and then the data are pushed to the data lakes. Organizations leverage the data lake for advanced analytics and insights to solve critical problems then.

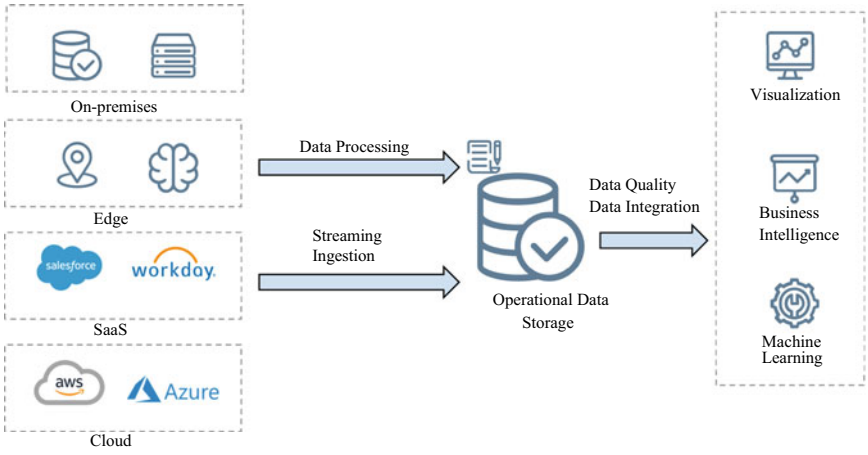

Fig. 4 Typical data processing framework

How can you create data processing pipeline architecture?

- Enterprise data preparation tools (EDPT) allow users to automatically convert data preparation steps into data pipelines that can be easily documented or modified. Traditional data preparation tools hinder this operationalized method of the data processing workflow.
- Traditional coding practices are using data processing technologies like Kafka, SQL, Spark and Pandas. These tools require understanding and expertise of the same so that they can be leveraged and used to streamline the data processing workflow

Data processing pipelines have been implemented for a lot of years now. Typical processes would be inclusive of reading data, transforming it in the required format and providing an output of a new data set for usage. Be it any use case from a wide array, a data processing pipeline must connect efficiently, collect effectively, integrate and prepare data as per requirement and deliver data at scale and also at the speed of the businesses at the same time.

5 Analytics and Insights

In a supply chain life cycle in the freight industry, a load is created when a demand wants to ship goods from point A to point B within x amount of time (where x can be any duration). Once the load is created, it gets exposed to the entire supply base or the fleet owners who own the trucks and are ready to provide trucks to ship the product. Once a supplier relates to the relevant demand, this is a successful matchmaking done. The supplier sends the truck to the loading point, and the products are loaded on the truck. In post loading, the truck is stated as truck in transit wherein the truck is moving from point A to point B. On reaching the unloading point, the products are delivered in the destination location, and the full and final settlement is done. This is how a supply chain looks like for an order in the freight businesses.

Now, these individual phases of the life cycle may be susceptible to a lot of issues and problems. Some of the typical issues might be

- Load expires even before the relevant supply could see it
- Document or truck-type mismatch resulting in order cancellations at the loading point
- Spillage or product damage discovered at unloading point leading to deductions
- Truck malfunction during the transit process
- Payment/fulfilment issues due to settlement document mismatches.

Hence, it becomes important to track each and every aspect of the supply chain life cycle, till the order is fulfilled to ensure smooth functioning of the same. This is where data capture comes into play. If data are tracked in each of these phases in the correct manner, advanced analytics methods and machine learning can be leveraged to solve these issues.

This is where companies use data to monitor for signals that are key to their businesses and solve for crucial problems.

6 Machine Learning Implementation

Once the data collection platforms and data analytics infrastructure are set up and ready to go, this is where organizations can leverage the rich data and implement and optimize supply–demand matchmaking, pricing and various other services KPIs [8]. Machine learning is a technique wherein data are used to understand the underlying pattern in the same. The algorithm learns the patterns from the data through iterations and helps in solving problems like forecasting, clustering, predictions and factor significance determination. Machine learning can broadly be classified into three categories, namely supervised, unsupervised and semi-supervised learning. In supervised learning, the input variables are known from the beginning wherein the algorithm is used to learn the mapping function between the input and the output variables. In unsupervised learning, we only have the input variables, and there is no understanding of the corresponding output variables. The goal of the algorithm is to understand the underlying structure and affinity of the data points with one another. Finally, semi-supervised are the types of problems with a large amount of input data out of which only some of the data are labelled. These algorithms/problems lie between supervised and unsupervised machine learning algorithms.

6.1 Demand–Supply Matchmaking

With the understanding of the regional requirements of fleet owners and customers in terms of supply and demand, the problems for deficit or surplus can be tackled for both parties. As shown in Fig. 6, the predicted demand has increased by the end of the year, whereas supply has remained constant even for the predicted duration. This seems to be a potential problem for the company as the suppliers remaining constant might lead to a deficit in supply base for the increasing demand. By understanding the predictions, organizations can ramp up on relevant requirements to avoid any form of issues. The models can account for unprecedented demand shocks and can help ramp up on requirements accordingly. Throughout the shipping life cycle, machine learning algorithms help organizations anticipate the complex dynamics of the system starting from tendering freight, matching trucks to load and hauling. This prediction of loads can be further split into other levels like truck-type mapping, regional demand–supply, cancellation affinity, trucker traits and many other factors (Fig. 5).

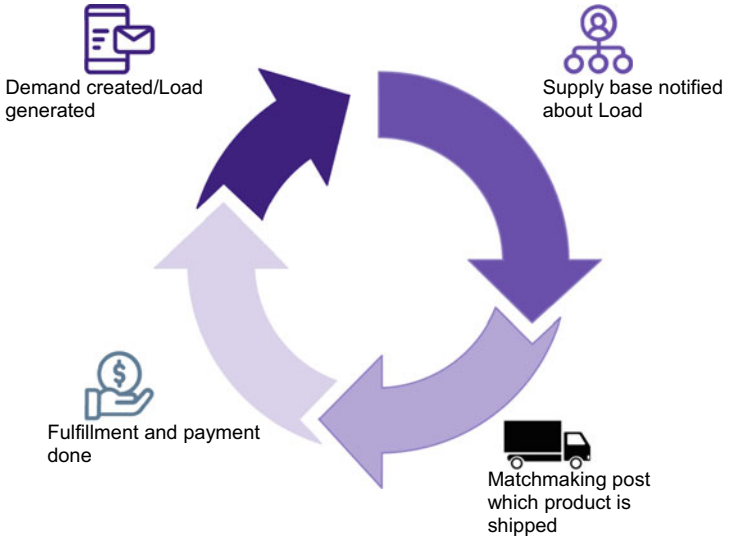

Fig. 5 Supply chain life cycle in freight

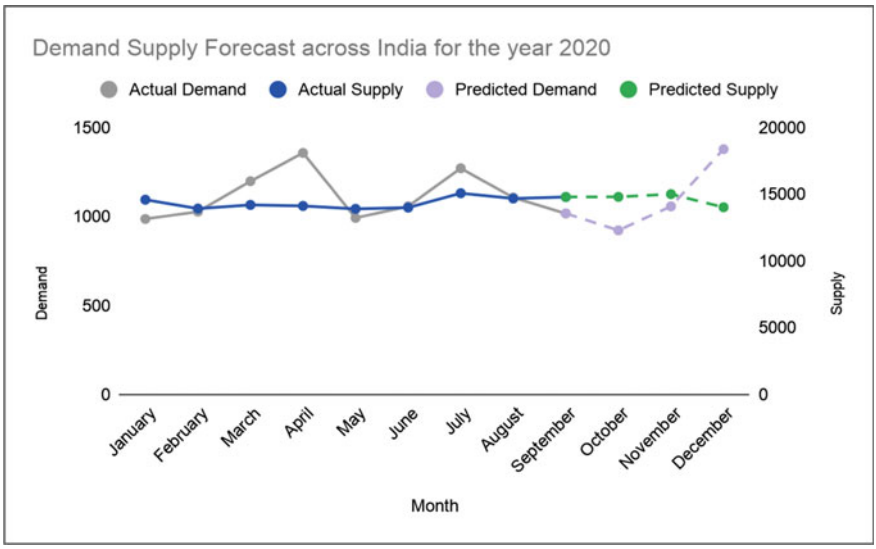

Fig. 6 Demand–supply predictions for fourth quarter of the year of trucking market

6.2 Pricing and Incentives

The freight market relies on the dynamic and complex nature of pricing to a big extent. Pricing especially in India is dependent on factors like market standard, seasonality,

affinity for products movement and a lot of other factors [9]. Without the efficient pricing model under play, there would be an enormous amount of cancellations and inefficient matchmaking processes. A typical instance would be if there is a load available to be sent from point A to point B, and it has been priced at $x per ton. But it has been understood that the price changed due to some seasonality factor or other extrinsic factor (maybe due to brokers negotiating at a cheaper price for the same load). The current market price is evaluated at $(x-50) per ton, and this would result in inefficient matches made for that load. This is where efficient pricing models come into play, wherein these extrinsic roles are accounted for while listing the appropriate price for the load. An effective model could account the factors well while stating the price, and it could also account for the dynamic nature of pricing in real time, while monitoring the same.

As depicted in Fig. 7, brokers intervene in the entire shipping cycle in terms of pricing, and this results in a revenue loss for the freight company. Hence, it is quite important to fulfil the pricing expectations for truckers in the market which can easily be captured by broker networks, especially in a broker heavy market of a country like India. It becomes important that these metrics and factors are kept in check in real time to avoid such bottlenecks for the companies. With effective data storage capacity and machine learning models, these daunting issues can be catered to.

Other than that, there is the active competitiveness of having the best incentive programs for drivers, consumers or truckers in any sort of businesses, who want to give the best that they can to their employees. There are companies which rely on user behaviour and transaction data to roll out the best schemes possible and incentivize their efforts. This in turn leads to maximization of effort and results and leads to retention of these drivers/truckers in their businesses. A typical example would be that of incentivizing the cab drivers (Ola/Uber) who are paid some bonuses for completing milestones on a daily cadence. This data on driver completions and incentive earnings can help organizations understand how affine these drivers are to stick to their business and how well they are bound to perform in the long run.

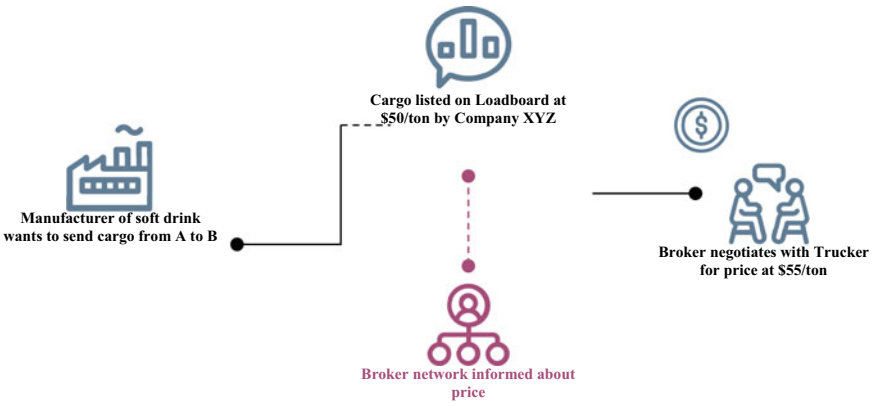

Fig. 7 Broker intervention for loads by gaming pricings

Added to that, these data sets could be fed into machine learning algorithms and the behaviour could be predicted, if their performances tend to go down. It is in the interest of the organizations that these drivers/truckers are incentivized well so that the businesses could be monetized to the maximum.

6.3 User Segmentations to Understand User Activities

Clustering is an unsupervised machine learning algorithm. In these types of problems, the user only has the input data, and the corresponding output variables are missing. The primary objective of the algorithm is to model the underlying structure of the data points in order to understand the data better. Unlike supervised machine learning algorithms, these algorithms do not rely on past data to understand the underlying pattern in the same. There is no correct answer as such, and hence, the algorithm is left on their own to discover the structure and underlying patterns in the data.

A lot of businesses use clustering techniques to understand their user base and their affinity towards the business. As shown in Fig. 8, a sample size of 5000 users were considered and various metrics like average number of trips, supply fleet size (or trucks owned) and average orders cancelled were used to define clusters of users. Cluster 1 seems to be the least engaging with the business while users belonging to Cluster 4 are the highly engaging users based on trips done, trucks owned and order cancelled. These overviews understanding of the user base can help roll out user-centric business plans, target users with custom notifications and other campaigns.

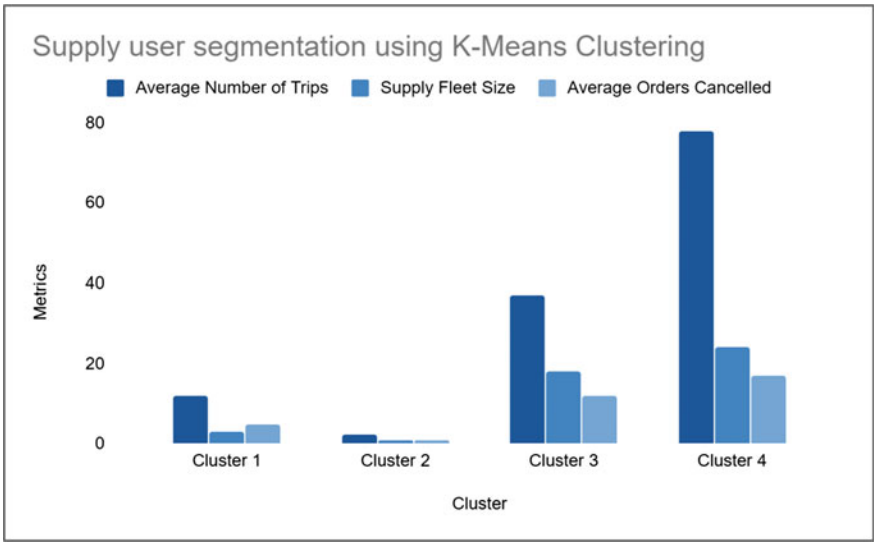

Fig. 8 Supply base segmentation based on user properties

K-means clustering was implemented for the activity with the number of clusters kept at 4. The number of clusters is identified using the elbow method for WSS calculation which provides the optimal number of clusters to be used while modelling.

Hence, clustering techniques can be very useful in a lot of industries ranging from freight to retail. Another instance of clustering implementation is in the retail sector wherein there is a need to understand customer buying patterns to be able to provide customer tailored offers and shelf placement optimization techniques. With proper clustering implementation, users can be segmented into different clusters like high value customers, returning customers, churned segments and highly active buckets. These buckets can then individually be targeted for marketing and other sorts of campaigns.

7 Comparative Study of Different Techniques

Even if we can categorize machine learning algorithms [10], the underlying assumptions and details that need to be considered while doing the same can be even more complex of a process. Below shown are some of the algorithms and their strengths and weaknesses pertaining to each of them. This section discusses which model to use and why you should use it, based on its pros and cons.

Algorithm	Strength	Weakness
Linear regression	Simple and the problem of overfitting can be avoided by regularization	Performs poorly when there are nonlinear relationships
Naive Bayes	They are easy to implement and can scale data over the time	Beaten by other models which are properly trained and tuned
K-means clustering	It is simple, fast and flexible if the data is pre-processed well, then significant features are engineered	The number of clusters must be determined before modelling, which is often not an easy task
Deep learning	Performs very well when classifying for text, audio and image data	It requires a large amount of data set to train due to which it is not considered as a general-purpose model
Classification trees	These algorithms are robust to outliers, and they can model nonlinear decision boundaries	Individual trees are very prone to overfitting
DBSCAN	This algorithm does not require each point to be assigned a cluster thereby reducing the noise of the clusters	The user must fine-tune the hyperparameters which define the density of the clusters

8 Chapter Takeaways and Significance

With the advancement of human requirements, the need to scale technology has also become the need of the hour. We have only discussed the freight industry in the chapter along with the implementation of machine learning and advanced analytics to solve problems. There are numerous other sectors which require the same level of business acumen and technological advancement to proceed ahead. Some of those industries being fintech, gaming, programmatic ads, health care and many more can leverage ML and advanced analytics to meet the consumer demands. This chapter focused on some of the key aspects of an analytics and machine learning infrastructure for a company dealing in the freight sector of the businesses.

We discussed data storage and big data ecosystems set-up in companies which are capable of processing terabytes of data in a couple of seconds. These types of processing power and infrastructure are required to solve problems and businesses in real time without having to suffer any down times. Even a single second of down time can incur huge losses on the consumer side of businesses, and hence, it becomes very important as to what type of data engineering or platforms you employ, who can set up this infrastructure up in the best optimized way. Another factor associated with data storage and data minimization efforts is the cost optimization aspect of it wherein organizations want to save dollar values worth of processing or storage systems. We also discussed what steps go into the data processing steps in the companies. Even before modelling or doing any sorts of predictions, it is quite important to understand the intricacies of the data and derive meaningful metrics out of them. These metrics in turn shape the overall outcome of the business along with conclusive predictions and model deployments. Finally, when the organization is all set with the data architecture and schemas set in place, along with the appropriate data pipelines, this is where the modelling phase starts and teams can start deploying real-time predictive or segmentation models for streamlined operations.

For further scope of research and improvements, these data storage technologies can be further researched upon to facilitate an even more streamlined process of capturing data and storing them in the right schema in the right manner. The optimal data processing tools need to be determined which can effectively result in cost and speed optimization for the teams. Other than that, deep learning can be implemented for incrementality in prediction accuracies.

9 Conclusion and Future Scope

With the current progression in the data science and AI domain and with the advent of efficient data collection techniques, there is wider scope to leveraging data to understand nuances in business problems. Real-time tracking mechanisms can be set up in the freight businesses to scale the supply chain. Machine learning and artificial intelligence tools can be used to predict demand in various geographical locations in

real time so that supply is available to the audience at the right time. These signals will pre-emptively be detected by the advanced algorithms so as to keep the supply chain engaged, thereby avoiding staleness in any indents or loads.

For instance, one of the most common issues faced in the freight industry is the expiry of indents, without prior engagements. In a typical supply chain process, an indent is created by demand. An indent is a load requirement that is created by the relevant demand team, who wants to ship some product from location A to location B. If this indent is not exposed to the relevant demand well within time, it might expire, thereby impacting the supply chain cycle. Advanced prediction mechanisms and signal detection techniques can track these demand creations in real time and expose them to the relevant supply base which would in turn deter indent expiration/staleness.

References

1. Facility Insights Report. 2020. Uber Freight. https://info.uberfreight.com/UberFreight-Facility-Ratings-Insight-Report_2020.html.
2. Tavasszy, Lóránt A. 2020. Predicting the effects of logistics innovations on freight systems: Directions for research. *Transport Policy* 86: A1–A6.
3. Ferrari, Paolo. 2014. The dynamics of modal split for freight transport. *Transportation Research Part E: Logistics and Transportation Review* 70: 163–176.
4. Harris, Gregory. 2017. A Freight Planning Framework. Research Issues in Freight Transportation, Congestion and System Performances, Number E-C225 November 2017, pp. 45–48.
5. Cruijssen, Frans, et al. 2007. Horizontal Cooperation in Transport and Logistics: A Literature Review. *Transportation Journal JSTOR* 46 (3): 22–39. www.jstor.org/stable/20713677. Accessed 25 Sept. 2020.
6. What is Data Lake? Amazon. https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/.
7. Data Processing Pipeline patterns. https://blogs.informatica.com/2019/08/20/data-processing-pipeline-patterns/.
8. How Machine Learning Improves the Efficiency of Freight Operations. https://www.freightwaves.com/news/how-machine-learning-improves-the-efficiency-of-freight-operations.
9. Freight Costs: An Insider's Look on Freight Pricing Buyers Should Know. https://www.intekfreight-logistics.com/freight-logistics-cost-buyers-guide.
10. Arun Kumar, P., S. Agrawal, K. Barua, M. Pandey, P. Shrivastava, and H. Mishra. 2020. Dynamic Rule-Based Approach for Shelf Placement Optimization Using Apriori Algorithm. In *Frontiers in Intelligent Computing: Theory and Applications. Advances in Intelligent Systems and Computing*, edited by S. Satapathy, V. Bhateja, B. Nguyen, N. Nguyen, D.N. Le, vol. 1014. Singapore: Springer. https://doi.org/https://doi.org/10.1007/978-981-13-9920-6_23.