Applying Data Mining Tools in Transportation: Data-Driven Supply Chain View

Sanjida Binte Islam¹, Md. Mamun Habib²

¹Information Systems and Operations Management, University of Nottingham, United Kingdom
²Industrial Engineering, University of Texas - Arlington (UTA), USA.

1aurin.sanjida@gmail.com
2mohammad.habib@uta.edu

Abstract— Despite the big data research and relevance of data analysis there has been limited empirical research and implication of data-driven supply chain networks. This paper explores the effect of data-driven supply chain capabilities on transportation (train based). In order to illustrate the shortest path calculation, London Underground Transportation open source data have been analysed through implementing different data mining tools and using programming language Python and R. The findings indicate that a data-driven supply chain has a significant time efficient effect on the logistics support. Coordination, using available data, and supply chain responsiveness are positively and significantly related to time and cost efficient performance. This system can be implemented in train based logistic support to consider the route selection.

Keywords— Big Data, Data-Driven Supply Chain, Logistics, Transportation, Data Mining, Analytics

1. Introduction

Transportation is one of the logistical drivers of supply chain performance. Deploying a big data strategy to the supply chain could potentially lead to improvements in efficiency and effectiveness through activities such as monitoring the location, transfer and acceptance of products and services, advanced demand forecasting and supply planning, and understanding behaviour of customers and suppliers [1]-[4]. In any case, the test is that distinctive store network individuals may utilize diverse data frameworks and innovations and be compelled to just access their own storehouses of information. To utilize the information to expand benefits, data should be shared across measures inside the association, yet additionally outside the association, accordingly giving a genuine start to finish measure view to all store network accomplices. In an information driven store network measure, data is shared across the whole inventory network to interface inventory network accomplices and give start to finish production network information access [5].

Almost a trillion dollars spent annually to the global economy on-road transportation [6]. Intelligent Integrated Transportation System which is also known as IITS is implemented in China to an integrated multimode system which is developed by using the available data of transportation, passenger flow efficiently [7]. The research suggested that a 100% database is growing in every twenty months [8]. Due to the exponential growth of data, analyzing and converting the data to essential resources are becoming unmanageable. However, only organizing the data can not provide the maximum output; that is why practical implements must be habituated to agnize, understand the data to turn into actionable information. Therefore, in recent times, data mining tools are gaining popularity. These tools convert the massive amount of unsupervised data into meaningful, usable data. It is a process of extracting and identifying paramount and subsequent cognizance from the immense database. It is a component of a process called knowledge discovery from databases [9]. Due to the availability of the enormous number of data in the train based transportation system, this is the topic of significant interest in big data analytics. Faulkner in 2002 first researched the systematic usage of data in railway transportation systems[10].

With the technological advancements across the entities of Supply Chain, data generated is increasing at a fast rate. The information flow was documented in terms of physical documents until the use of Information Technology in Supply Chain. Majority of the information flow linked to the material flow is being documented in the form of digital structured data. As the scope of Supply Chain is currently worldwide, the volume of data collected from its numerous processes and the velocity at which it is being generated can be qualified as Big Data. Recently professionals from the supply chain department collect and analyze data that recommends new techniques of organizing and support decisions quantitatively [11]. Big data refers
to data that includes huge volume, velocity, and variety that need proper systems to process it [12],[13]. In the context of supply chains, the use of big data as the basis for quantitative and qualitative techniques aimed at improving supply chain competitiveness. It will give organizations an edge to minimize risk and cost. While data is growing in importance as a driver of better decision support and improved business performance for those firms able to leverage it [14], it has been suggested that not all firms are able to translate investments in computational infrastructure into performance gains [15].

4V’s of Big Data Analytics velocity, volume, variety, veracity are not the only concern, but the fifth V engendering value out of a substantial amount of data. With the congruous application of data analytics, it is possible[16].

Figure 1. 4V of Big Data Analytics (Source: Authors)

2. Literature Review

Accomplishing supply chain network viability and productivity enhancements expects admittance to information from various utilitarian territories of an association and from various production network accomplices. Supply Chain Management incorporates utilizing both interior and outside assets/data to encourage store network exercises [17]-[19]. This can be utilized to acquire an upper hand. Explicitly to deliver customer/market necessities and work with exchanging accomplices to make request winning items and administrations. We conceptualize SCC as a multidimensional build that incorporates four measurements: data trade, coordination, bury firm action combination, and store network responsiveness. Every one of the four measurements mirrors a capacity to cooperatively perform cross-practical, e.g., joint effort across item/administration configuration, buying, creation, deals/showcasing, and dissemination works, and between hierarchical exercises, e.g vital data sharing and coordination between a central firm and its inventory network accomplices, that are needed in the store network measure [20],[21].

Data trade alludes to the capacity of a firm to deliberately share information/data about items and favorable to cress with its production network accomplices in a compelling and effective way. Earlier examination has uncovered information trade to be a significant store network ability. Completely created, it can empower a firm to accomplish successful and proficient progressions of items and administrations, data and pull away from contenders . For instance, such a combination has appeared to assist the firm with creating creation designs and convey items and administrations on schedule.

Hu et al. have proposed a decision tree model to optimize pricing. The developed model is established in Taiwan . Sun et al. have developed a neural network between waiting time of passengers, the demand of passengers, time of arrival and departure of the train [22]. This research proposed a train scheduling system which reduces waiting time of the passengers. Another similar work provides a model by applying a neural network by utilizing historical data of passengers to forecast the short-term passenger demand [23].

Authors have utilized linear regression varying train advent time depending on historical data of the station [24]. They proposed this model to estimate the delay of the advent of the train. In another research authors did quantitative research on data accumulated from the keenly intellective cards to estimate routes cull pattern of passengers [25]. They utilized the probabilistic model to determine the number of passengers for each train analyzing historical data [26]. Following this research, another research proposed implementation of a support vector machine to evaluate the accuracy of a previously proposed system [27]. Authors suggested to use mean squared error and mean average percentage error to validate the prediction. They have proposed a predictive control model based on regulations of the train. They implied that uncertain passenger flow affects the control law. Using MATLAB authors showed how the waiting time of the train, passenger flow affect the scheduling of a metro system. In another research authors surveyed passengers to monitor their reaction during an emergency [28]. According to results, authors suggested injunctive authorization for the station management [29]. They used relegation implements for the suggestion.

Authors have analyzed passenger flow from tube to bicycle utilizing available TFL data and oyster data and identified the co-cognition between bicycle usage and passenger flow at busy stations by
implementing linear regression analysis tools [30]. In another research authors have used clustering techniques to identify a pattern between traveller’s journey and passenger flow at the station [31]. Then developed a prediction model using time series analysis and evaluated the results implementing MAPE. Researchers have proposed a prediction model by applying neural network technique between modes of transportation passenger use to reach stations, traffic conditions and number of passengers exit, entering stations [32]. Authors proposed a predictive model by using exponential smoothing to forecast bus advent time by considering each step between 1-minute time zones [33]. Authors proposed a data mining tool to relegate the track irregularity and to use smoothing exponential predicted next deviation when may transpire in railway track [34].

3. Methodology

London underground tube has been used as a contextual analysis to demonstrate the proposed data mining tools. The secondary data is collected from open data source Transportation for London (TFL). The data is available for the year 2017. The research purpose is to use available data of different station’s lines, zone, rail, latitude and longitude.

To calculate and recommend the shortest route between two stations. This system would be valid for any Railway Transportation to improve planning, improvising resources. In this paper four steps are implemented, three step-Preparation of data e.g., data processing, training and second step is simulation and validation; the last step is route suggestion. An overview is shown in figure 2. It describes the steps of process and how it concludes a solution for the next step.

In order to achieve the aim of the research, HITS algorithm has been utilized. Each of these algorithms has its own features that add to our study’s objectives. Frameworks begin with historical data which are collected from different sources. For data processing, we need a scalable and effective mechanism to convert an immense amount of raw data into supervised data. All the information can be stored in a cloud.

Figure 2. Method for the Research (Source: Authors)

Four distinct algorithms have been utilized, which are Holt-Winters smoothing exponential, Wilcoxon signed-rank test, be-spoke test to rank the station. Each of these algorithms has its own features that add to our study’s objectives. For data processing, we need a scalable and effective mechanism to convert an immense amount of raw data into supervised data.

To rank the station according to it’s traffic flows, these algorithms have been implemented. Smoothing exponential is used for prediction purposes in this research. Three parameters in smoothing exponential, we can use Holt-Winters. Keeping three parameters make this system compatible with any railway transportation system. Like three parameters alpha, beta, gamma could be utilized for level, trend, seasonality coefficient. For this research from TFL data seasonality is not available; therefore, gamma is kept as FALSE. The prediction interval is set to TRUE to define corresponding index attributes. Different coefficient values less than 1 of \[\alpha\] and \[\beta\] have been tried to get optimized results. Therefore, the alpha value is defined by 0.2 and beta 0.1 to get well-fitted results.

To check the performance of the predictive model, we have simulated data. For the testing process, we have executed the system R-times(e.g., R=1,000 iterations) so that it could examine the consistency of predicted values. This also enables us to identify the average precision or error values by combining all the models’ outcomes.

As shown in figure 2, the set.seed in R generates a particular set of random values. Set. Seed has been set at 100 because a fixed number obtain the same results given the same sequence every time for random values. For data simulation, be-spoke t-test has been implemented. For be-spoke test to generate data for each station lowest average passenger count from dataset has been defined as lower boundary and highest average passenger count as upper boundary for generating random passenger. It will generate...
random value between these boundaries. Now, if the chosen number is greater than equal to the average of predicted data we consider it as correct value, otherwise incorrect. Using Wilcoxon signed-rank test accuracy of data is evaluated.

For route selection the dataset has been prepared, which includes stations, lines, zone, rail, latitude and longitude. Two different datasets of station description and number of passengers travelling in stations are initially merged. A column of busy_rank has been added. Based on the number of passengers traveling, all the stations were given rankings from 1 to 10. In order to calculate the weight of busy first we ordered the stations based on the average passenger travelling descendingly. Then we pick a max value from the order and divide it up by 10. While setting the busy weight of each station we subtract from the max value from the order divided by 10. In a way clustering the weights of stations on average passenger count has been done. Thus, the greater the value of busy_weight column is, the larger the bottleneck is.

The proposed system has been kept as dynamic so that it can be implemented in any other railway based transportation system. These datasets can also be utilized to recommend routes, schedule. Station connections and number of passengers are utilized to rank the busiest stations. Based on passenger numbers implementing the HITS algorithm, we could get the weight of every station. This phase would help us to visualize the condition on the map. Bokeh library in python has been used to get the visualization. This is the final phase of the system. Busiest station monetization leads to shortest path calculation between stations.

Flowchart of the entire system has been discussed. Proposed work is a smart system which is integrated with data mining tools, Holt-winter’s prediction model, accuracy measurement, computational tools and shortest route suggestion.

![Flowchart of the System](source: Authors)

4. Discussion

Holt-Winters smoothing exponential model has been used for prediction. The system predicts the average daily passenger number. A system we have set for thirty days ahead; therefore, it predicts for 30 days. We can change it according to the requirement. We used the data generated by Holt-Winters method on the total number of passengers in a station. Then, we pass the value to predict function which generates predictions and bounds with column names fit, lower and upper. At below figure only one station Kings Cross St Pancras predicted values has been shown.
At figure 4, thirty predicted values are observed with upper and lower boundaries. Excluded from boundaries, the value will be outliers. It included the fitted value.

4.1 Evaluation of Prediction

By implementing a bespoke test, we have generated future data and compared predicted value with random data using Wilcoxon signed-rank test. It assumes that the distributions are not known, but that the parameters are not included. If the median is the same then it is considered as zero hypothesis. If the p-value is greater than 0.05 it would be a null hypothesis. If we reject the null, then the distribution is required to be checked whether it’s moving right or left of the median. We computed the accuracy rate of predicted value. We ran 1000 iterations and each time picked a random number between the minimum average and maximum average of a station. We considered random data to be correct if the average number of passengers in predicted results is greater than or equal to it. The accuracy rate of the predicted data is 96%. By applying this method, all the predicted values can be verified.

4.2 ARIMA Model

ARIMA is a predictive statistical method for time series. ARIMA is an abbreviation for the embedded auto-regressive-moving average. It is a design category which includes a sequence of typical time series structures. The three orders parameters of an ARIMA model, (seasonality p, trend d, noise q). The well-trained model of ARIMA describes the time series and is often in calculation we observe AIC, BIC which are Akaike's Information Criterion, the Bayesian Information Criterion. We have explicitly disabled warning messages to prevent a warning excess of warning messages, so that certain ARIMA parameter combinations may lead to numerical specifications. These mistakes can also lead to mistakes and an exception, so we ensure that we take these exceptions and overlook the combinations of parameters that trigger the problems. We have recognized a set of parameters using grid search, which generates the models which best match our time series data. This specific model can be analyzed in more detail. Therefore, ARIMA (1,1,2) and ARIMA(2,1,3) have been applied.

4.3 Implementing ARIMA (1, 1, 2)

To implement ARIMA(1, 1, 2), we have used the same dataset as smoothing exponential. We have used optimal parameter 1,1,2 for seasonality, trend and noise. The true argument guarantees that we generate one-step forecasting, which means that forecasts are produced at every stage using the entire history. In R 30 days, the prediction has been set.

4.4 Assess the result of ARIMA (2,1,3)

After running the ARIMA method in R, we have got the below-observed results.
4.5 Assess the result of ARIMA (1,1,2)

ARIMA (2,1,3) is highly parameterized; therefore, it did not fit the trend. We have tried ARIMA (1,1,2) for better results. The model precision can be evaluated with accuracy once the model has been produced. The Accuracy feature returns a MASE value to evaluate the model's precision. The best model is selected from the following outcomes, which show comparatively low ME, RMSE, MAE, MPE, MAPE, MASE values.

4.6 Comparing ARIMA and Holt Winters smoothing Exponential

From a single iteration, ARIMA calculates least-squares. We cannot vary seasonality in ARIMA that is why Holt Winters smoothing exponential process is more accurate. Comparing both results of the ARIMA method ARIMA (1,1,2) gives approximately the same results as Holt Winters. It is observed that both of the processes give the same trends, but later, one is much more efficient. That is why we have implemented smoothing exponential.

5 Recommendation

Stations have been weighted based on the number of passengers. Generated ranks of stations are tested with the HITS algorithm. Shortest path is calculated in two ways. By implementing prediction and shortest path calculation overview of the network's requirements of route planning, staff planning can be understood. In several research papers how implementing data analytics tools have changed operations planning of transportation are discussed [35].

Figure 6. Prediction Results of King's Cross Station in ARIMA(2,1,3) (Source: Authors)

From figure 6 it can be observed that in auto best-fitted parameter option R picked ARIMA(2,1,3) for the prediction. Log-likelihood discovers parameter values that maximize the chance of getting the information we observed. It is shown above model parameters. R calculates log probability for parameters. The sign for the MA portion is in accordance with this formula. Additional data is also printed along with model parameters.

Figure 7. Prediction Results of King's Cross Station in ARIMA(1,1,2) (Source: Authors)

We see that more than two standard variances from zero are observed in all parameters. Thus, the t-test passed all parameters. The model also calculated the value of the model's error term. The probability and aic values are also given. The ARIMA(1,1,2) model is best based on probability and aic. The third model is best based on likelihood and aic.
In this paper, it is not implying that this is the first proposed approach of prediction or shortest path calculation; instead, it is the first proposal which focuses on alternate shortest route calculation between two stations based on several attributes.

An essential aspect of the work presented in this research is routes between two stations and prediction using data mining, statistical tools and several algorithms to test proposed systems.

5.1 Future Work

For future work passenger flow, the available commute near stations could be considered. From train based transportation system data, it can be stored in the cloud. After normalizing the data with clustering, classification patterns can be identified.

As shown in figure 8, the dataset required to be prepared compatible for prediction and route planning. As observed implementing Holt-Winters smoothing exponential the passenger flow of train based transportation can be predicted. Minitab can be utilized for this. It has options to calculate residuals, lower and upper prediction limits. As observed in this paper, stations can be weighted based on the number of passengers, traffic flow. Depending on the weight of the station the shortest routes can be suggested and visualized. Proposed future work, recommends smart integration of logistics support with a view to make a cost-effective and time favourable plan.

6 Conclusion

In this paper an approach towards the shortest route between two stations by implementing has been shown considering and predicting the future of busy station ranking. Rankings would be changed depending on the traffic flow season. Two shortest paths can be generated between two stations. This system can be applied to any other railway based transportation. The London Underground tube data is utilized as a case study to demonstrate the effectiveness of the proposed approach. We ranked the stations and proposed alternate shortest paths based on the implementation of the weighted algorithm for stations and time. We also examined passenger flows and the place of the stations, such as the impact of stations. All the results have been evaluated by applying different methods. Four approaches and two programming languages are taken into account to identify the patterns of the dataset, prepare datasets, develop prediction models and suggest the shortest route selection.

The framework can be applied to any train based transportation before choosing any station to transport goods. The parameters will consider different parameters and suggest route. There is plenty of unstructured data available at train based transportation and analysts are utilizing these to come up with new solutions for transportation. In this paper it is suggested to utilize a large volume of available data as an external factor of Supply Chain Network Logistics support.

References


