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

Sanjida Binte Islam[#1], Md. Mamun Habib[*2]

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

[1]aurin.sanjida@gmail.com
[2]mohammad.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 effective- ness 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

International Journal of Supply Chain Management
IJSCM, ISSN: 2050-7399 (Online), 2051-3771 (Print)

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].

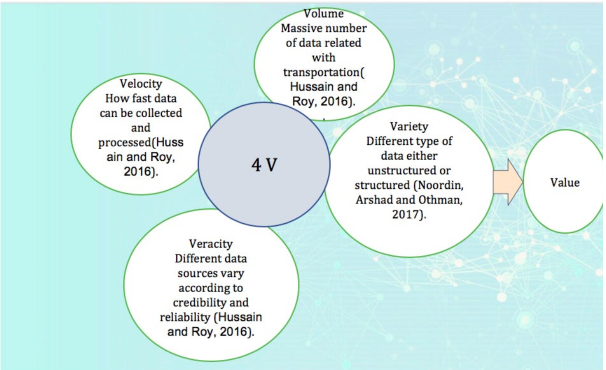

Figure 1. 4V of Big Data Analytics (Source: Authors)

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 cess 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-cognation 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 stop 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.

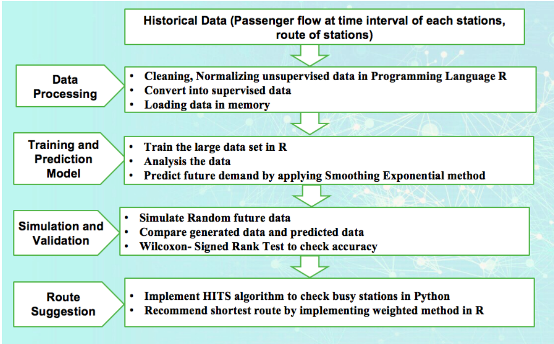

Figure 2. Method for the Research (Source: Authors)

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 α and βhave 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.

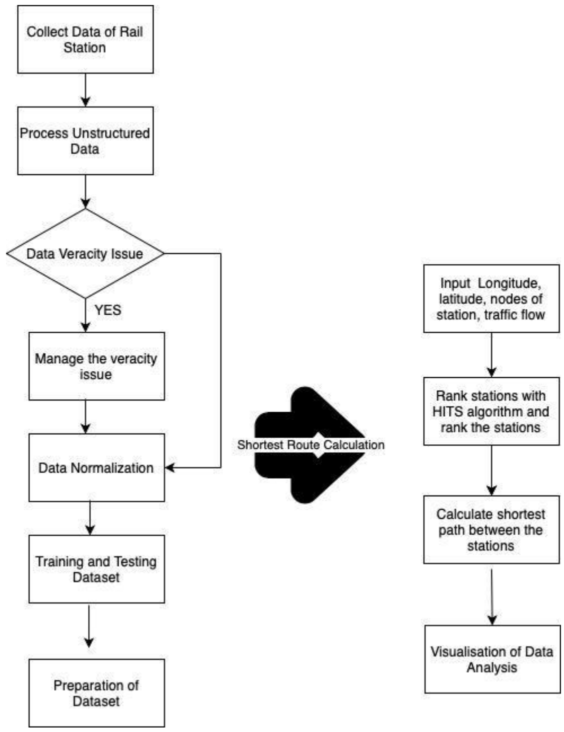

Figure 3. Flowchart of the System (Source: Authors)

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.

```
> forecasting_passenger_by_station("Kings Cross St Pancras")
Time Series:
Start = 163
End = 192
Frequency = 1
          fit      upr        lwr
163 1022.0120 10599.33  -8555.308
164 1007.0995 10813.45  -8799.253
165  992.1870 11064.32  -9079.948
166  977.2745 11352.65  -9398.106
167  962.3621 11678.69  -9753.969
168  947.4496 12042.27 -10147.370
169  932.5371 12442.87 -10577.791
170  917.6247 12879.69 -11044.444
171  902.7122 13351.75 -11546.330
172  887.7997 13857.91 -12082.310
173  872.8872 14396.93 -12651.156
174  857.9748 14967.55 -13251.598
175  843.0623 15568.48 -13882.358
176  828.1498 16198.48 -14542.180
177  813.2373 16856.32 -15229.845
178  798.3249 17540.84 -15944.188
179  783.4124 18250.93 -16684.106
180  768.4999 18985.56 -17448.559
181  753.5875 19743.75 -18236.573
182  738.6750 20524.59 -19047.242
183  723.7625 21327.25 -19879.721
184  708.8500 22150.93 -20733.226
185  693.9376 22994.91 -21607.031
186  679.0251 23858.51 -22500.463
187  664.1126 24741.12 -23412.895
188  649.2001 25642.15 -24343.751
189  634.2877 26561.07 -25292.491
190  619.3752 27497.37 -26258.617
191  604.4627 28450.59 -27241.666
192  589.5503 29420.30 -28241.204
```

Figure 4. Prediction Table for KingsCross St Pancras Station (Source : Authors)

At figure 4, thirty predicted values are observed with upper and lower boundaries. Excluded from boundaries, the value will be outliners. It included the fitted value.

4.1 Evaluation of Prediction

By implementing a be-spoke 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.

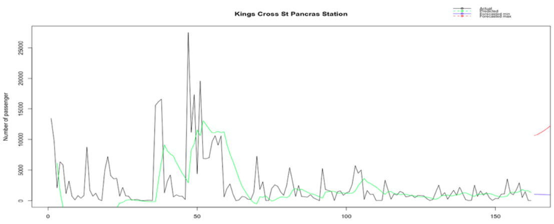

Figure 5. Comparing actual data with predicted value (Source: Authors)

In figure 5, we observed the trend of actual data and predicted value. The predicted value is represented as green, actual data as black, if we forecast further maximum value could go towards the red line and minimum value could go towards the blue line. We can observe that for King's Cross Station minimum passenger flow could be around 100.

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.

```
> forecast_passenger_with_arima("Kings Cross St Pancras")
Forecasted Data
    Point Forecast     Lo 80    Hi 80     Lo 95     Hi 95
163       71.40199 -4382.747 4525.551 -6740.631  6883.435
164      807.63478 -4136.990 5752.260 -6754.516  8369.786
165      849.58184 -4505.746 6204.909 -7340.685  9039.849
166     1018.83108 -4342.713 6380.375 -7180.943  9218.605
167      963.07950 -4480.285 6406.444 -7361.828  9287.987
168     1027.61655 -4420.162 6475.395 -7304.041  9359.275
169      988.26215 -4501.036 6477.560 -7406.894  9383.419
170     1019.96292 -4478.300 6518.226 -7388.905  9428.831
171      997.61316 -4530.077 6525.303 -7456.259  9451.485
172     1014.36777 -4526.271 6555.007 -7459.308  9488.044
173     1002.16445 -4562.631 6566.960 -7508.456  9512.785
174     1011.17287 -4569.087 6591.433 -7523.098  9545.444
175     1004.56448 -4597.219 6606.348 -7562.624  9571.753
176     1009.42646 -4609.229 6628.082 -7583.565  9602.418
177     1005.85425 -4632.917 6644.626 -7617.902  9629.611
178     1008.48051 -4647.893 6664.854 -7642.197  9659.158
179     1006.55027 -4669.145 6682.246 -7673.677  9686.778
180     1007.96915 -4685.675 6701.613 -7699.708  9715.646
181     1006.92624 -4705.564 6719.416 -7729.573  9743.426
182     1007.69283 -4722.880 6738.265 -7756.462  9771.847
183     1007.12935 -4741.982 6756.240 -7785.377  9799.636
184     1007.54353 -4759.670 6774.757 -7812.648  9827.735
185     1007.23909 -4778.296 6792.774 -7840.973  9855.451
186     1007.46287 -4796.133 6811.059 -7868.371  9883.297
187     1007.29838 -4814.451 6829.048 -7896.299  9910.896
188     1007.41929 -4832.318 6847.157 -7923.689  9938.527
189     1007.33042 -4850.420 6865.081 -7951.326  9965.986
190     1007.39574 -4868.254 6883.045 -7978.635  9993.426
191     1007.34773 -4886.189 6900.885 -8006.039 10020.735
192     1007.38302 -4903.956 6918.722 -8033.230 10047.996
Series: ts_data
ARIMA(2,1,3)

Coefficients:
          ar1     ar2      ma1      ma2      ma3
      -0.3919  0.2522  -0.1260  -0.4755  -0.2811
s.e.   0.4547  0.3791   0.4927   0.1083   0.3308

sigma^2 estimated as 12079732:  log likelihood=-1539.43
AIC=3090.86   AICc=3091.4   BIC=3109.35

Training set error measures:
                    ME     RMSE      MAE       MPE    MAPE     MASE        ACF1
Training set -365.9321 3410.621 2008.209 -2305.849 2332.08 1.006521 -0.01912311
                    ME     RMSE      MAE       MPE    MAPE     MASE        ACF1
Training set -365.9321 3410.621 2008.209 -2305.849 2332.08 1.006521 -0.01912311
                    ME     RMSE      MAE       MPE    MAPE     MASE        ACF1
Training set -365.9321 3410.621 2008.209 -2305.849 2332.08 1.006521 -0.01912311
```

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.

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.

```
>
>
> forecast_passenger_with_arima("Kings Cross St Pancras")
Forecasted Data

Forecast method: ARIMA(1,1,2)

Model Information:

Call:
arima(x = ts_data, order = c(1, 1, 2))

Coefficients:
         ar1      ma1     ma2
      0.7228  -1.3005  0.3199
s.e.  0.1112   0.1354  0.1261

sigma^2 estimated as 12551684:  log likelihood = -1545.01,  aic = 3098.03

Error measures:
                   ME     RMSE      MAE       MPE     MAPE     MASE        ACF1
Training set -434.844 3531.884 2022.479 -2434.611 2453.463 1.013673 -0.03934095

Forecasts:
    Point Forecast     Lo 80    Hi 80     Lo 95     Hi 95
163       631.6821 -3908.652 5172.016 -6312.159  7575.524
164       814.2471 -4114.356 5742.850 -6723.400  8351.895
165       946.2010 -4198.098 6090.500 -6921.326  8813.728
166      1041.5743 -4230.467 6313.615 -7021.316  9104.465
167      1110.5080 -4241.525 6462.541 -7074.721  9295.737
168      1160.3317 -4244.615 6565.278 -7105.820  9426.484
169      1196.3432 -4245.594 6638.280 -7126.381  9519.067
170      1222.3715 -4246.883 6691.626 -7142.132  9586.875
171      1241.1842 -4249.334 6731.702 -7155.838  9638.207
172      1254.7816 -4253.103 6762.666 -7168.801  9678.364
173      1264.6095 -4258.073 6787.292 -7181.605  9710.824
174      1271.7128 -4264.038 6807.463 -7194.487  9737.913
175      1276.8470 -4270.785 6824.479 -7207.524  9761.218
176      1280.5579 -4278.131 6839.246 -7220.723  9781.838
177      1283.2400 -4285.923 6852.403 -7234.060  9800.540
178      1285.1786 -4294.044 6864.401 -7247.507  9817.864
179      1286.5797 -4302.405 6875.564 -7261.035  9834.194
180      1287.5924 -4310.937 6886.122 -7274.619  9849.804
181      1288.3244 -4319.591 6896.240 -7288.242  9864.891
182      1288.8535 -4328.330 6906.037 -7301.887  9879.594
183      1289.2359 -4337.126 6915.598 -7315.543  9894.014
184      1289.5123 -4345.961 6924.986 -7329.201  9908.225
185      1289.7120 -4354.820 6934.244 -7342.854  9922.278
186      1289.8564 -4363.692 6943.405 -7356.500  9936.212
187      1289.9608 -4372.570 6952.491 -7370.133  9950.054
188      1290.0362 -4381.448 6961.521 -7383.751  9963.823
189      1290.0907 -4390.323 6970.505 -7397.353  9977.534
190      1290.1301 -4399.192 6979.452 -7410.937  9991.198
191      1290.1586 -4408.052 6988.370 -7424.503 10004.820
192      1290.1792 -4416.903 6997.261 -7438.050 10018.408
```

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.

4.6 Comparing ARIMA and Holt Winters smoothing Exponential

From a single iteration, ARIMA calculates least-squares. We can not 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].

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.

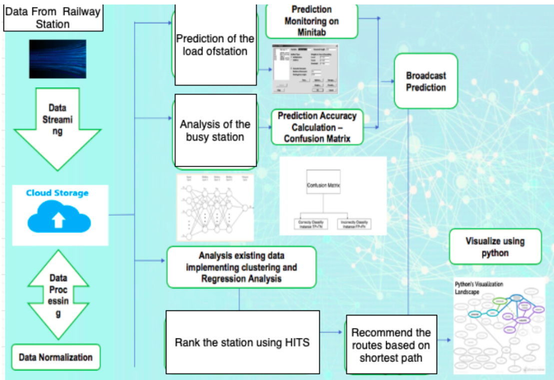

Figure 8. Recommended Future Work (Source: Authors)

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

[1] Davenport, T., 2006. *Competing on Analytics*. [online] Harvard Business Review. Available at: <https://hbr.org/2006/01/competing-on-analytics> [Accessed 1 February 2021].

[2] Davenport, T.H., Barth, P., Bean, R., 2012. How ''big data'' is different. MIT Sloan Manage. Rev. 54 (1), 43–46.

[3] Kwon, O., Lee, N., Shin, B., 2014. Data quality management, data usage experience and acquisition intention of big data analytics. Int. J. Inf. Manage. 34 (3), 387–394.

[4] Waller, M., Fawcett, S., 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. J. Bus.Logist. 34, 77–84.

[5] Sanders, N.R., 2014. Big Data Driven Supply Chain Management: A Framework for Implementing Analytics and Turning Information Into Intelligence.Pearson Financial Times.

[6] El Hatri, C. and Boumhidi, J. (2017). Traffic management model for vehicle re-routing and traffic light control based on Multi-Objective Particle Swarm Optimization. *Intelligent Decision Technologies*, 11(2), pp.199-208.

[7] Hosu, A., Varga, M., Kiss, Z., Polgar, Z. and Ivanciu, I. (2015). Integrated ubiquitous connectivity and centralised information platform for intelligent public transportation systems. *IET Intelligent Transport Systems*, 9(6), pp.573-581.

[8] Salavati, A., Haghshenas, H., Ghadirifaraz, B., Laghaei, J. and Eftekhari, G. (2016). Applying AHP and Clustering Approaches for Public Transportation Decisionmaking: A Case Study of Isfahan City. *Journal of Public Transportation*, 19(4), pp.38-55.

[9] Gothane, S. (2011). Predictive Analysis In Data Mining Using Weighted Associative Classifier. *Indian Journal of Applied Research*, 1(6), pp.115-119.

[10] Faulkner, A. (2002). *Safer Data: The use of data in the context of a railway control system*. 1st ed. London: Springer, London, pp.217-230.

[11] Hazen, B.T., Boone, C.A., Ezell, J.D., Jones-Farmer, L.A., 2014. Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications. Int. J. Prod. Econ. 154, 72–80.

[12] Lycett, M., 2013. 'Datafication': making sense of (big) data in a complex world. Eur. J. Inform. Syst. 22, 381–386.

[13] MacCallum, R.C., Browne, M.W., Sugawara, H.M., 1996. Power analysis and determination of sample size for covariance structure modelling. Psychol. Methods 1 (2), 130–149.

[14] Stank, T., Crum, M., Arango, M., 1999. Benefits of interfirm coordination in food industry supply chains. J. Bus. Logist. 20 (2), 21–41

[15] Taylor, D., 2003. Supply chain vs. supply chain. Computerworld 37 (45), 44–45. The Economist, 2010. Data, Data Everywhere. Special Report on Managing Information.

[16] Hongying, X. (2017). The Application of Database Technology in Information Society and its Existing Problems. *Big Data and Cloud Innovation*, 1(1).

[17] Amit, R., Schoemaker, P.J., 1993. Strategic assets and organisational rent. Strateg. Manage. J. 14 (1), 33–46.

[18] Collis, D.J., 1994. Research note: how valuable are organizational capabilities? Strateg. Manage. J. 15 (8), 143–152

[19] Wu, F., Yeniyurt, S., Kim, D., Cavusgil, S.T., 2006. The impact of information technology on supply chain capabilities and firm performance: a resource-based view. Ind. Mark. Manage. 35 (4), 493–504.

[20] Wong, C.Y., Boon-itt, S., Wong, C.W.Y., 2011. The contingency effects of environmental uncertainty on the relationship between supply chain integration and operational performance. J. Operat. Manage. 29, 604–615.

[21] Zhao, X., Huo, B., Selend, W., Yeung, J.H.Y., 2011. The impact of internal integration and relationship commitment on external integration. J. Operat. Manage. 29, 17–32.

[22] Hu, Z., Yan, Y. and Qiu, Z. (2008). Research on Optimization Model of Making Inter-city Passenger Train Operation Plan and Ticket Price. *n Proceedings of the International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII'08)*, 3, pp.45–48.

[23] Xie, M., Li, X., Zhou, W. and Fu, Y. (2014). Forecasting the Short-Term Passenger Flow on High-Speed Railway with Neural Networks. *Computational Intelligence and Neuroscience*, 2014, pp.1-8.

[24] Wang, R. and Work, D. (2015). Data driven approaches for passenger train delay estimation. *n Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC), Las Palmas, Spain*, pp.535–540.

[25] Zhao, J., Zhang, F., Tu, L., Xu, C., Shen, D., Tian, C., Li, X. and Li, Z. (2017). Estimation of Passenger Route Choice Pattern Using Smart Card Data for Complex Metro Systems. *IEEE Transactions on Intelligent Transportation Systems*, 18(4), pp.790-801.

[26] Abadi, A. and Wutsqa, D. (2014). Neuro fuzzy model with singular value decomposition for forecasting the number of train passengers in Yogyakarta. In: *n Proceedings of the 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. China, pp.178–182.

[27] Wang, P., Wu, C. and Gao, X. (2016). Research on subway passenger flow combination prediction model based on RBF neural networks and LSSVM. In: *In Proceedings of the 2016 Chinese Control and*

Decision Conference (CCDC). IEEE, pp.6064–6068.

[28] Li, S., De Schutter, B., Yang, L. and Gao, Z. (2016). Robust Model Predictive Control for Train Regulation in Underground Railway Transportation. *IEEE Transactions on Control Systems Technology*, 24(3), pp.1075-1083.

[29] Shiwakoti, N., Tay, R., Stasinopoulos, P. and Woolley, P. (2017). Likely behaviours of passengers under emergency evacuation in train station. *Safety Science*, 91, pp.40-48.

[30] Aslam, N. and Cheng, T. (2015). Big Data Analysis of Population Flow between TfL Oyster and Bicycle Hire Networks in London. London: University College London.

[31] Lathia, N., Froehlic, J. and Capra, L. (2010). Public Transport Usage For Personalised Intelligent Transport Systems. *2010 IEEE International Conference on Data Mining Mining*, (11770455).

[32] Aqib, M., Mehmood, R., Alzahrani, A., Katib, I., Albeshri, A. and Altowaijri, S. (2019). Rapid Transit Systems: Smarter Urban Planning Using Big Data, In-Memory Computing, Deep Learning, and GPUs. *Sustainability*, 11(10), pp.27-36.

[33] Xiong, G., Liu, Z., Liu, X., Zhu, F. and Shen, D. (2013). *Service science, management, and engineering*. 2nd ed. Woodhead Publishing, pp.117-140.

[34] Zhao, J., Zhang, F., Tu, L., Xu, C., Shen, D., Tian, C., Li, X. and Li, Z. (2017). Estimation of Passenger Route Choice Pattern Using Smart Card Data for Complex Metro Systems. *IEEE Transactions on Intelligent Transportation Systems*, 18(4), pp.790-801.

[35] Baru, C., Bhandarkar, M., Nambiar, R., Poess, M. and Rabl, T. (2013). Benchmarking Big Data Systems and the BigData Top100 List. *Big Data*, 1(1), pp.60-64.