Air Passenger Traffic Forecasting using Community detection and LSTM

Gayathri Sridhar
Dublin City University
School of Computing
Dublin, Ireland

Student ID: 20211232
gayathri.sridhar2@mail.dcu.ie

Meenakshi Srinivasan
Dublin City University
School of Computing
Dublin, Ireland

Student ID: 20210360
meenakshi.srinivasan2@mail.dcu.ie

***Abstract*—The Aviation industry has seen a significant growth in the recent years and has contributed to the upturn in the global economy. Air passengers forecasting[1] has become inevitable for the operating airlines to plan their flights to cope up with the growing demands and the passengers can also plan their journey beforehand. The main objective of this paper is to predict the trends in the number of passengers travelling all over the world. As the travel trends differ from season to season and certain airports are busier than others, the predictions would be more specific, if it is based on seasons and hubs. In order to identify strongly connected routes and city pairs in the air transportation network, the Louvain algorithm[6] for community detection is used. The busiest hubs are identified using this algorithm and then, season-wise forecasting is implemented for air passengers traffic. Long Short Term Memory(LSTM)[5] model along with k-fold cross validation is used for the prediction purpose. The performance of the LSTM model is then evaluated using standard error metrics [19] such as Mean Absolute Error(MAE),R^2 score and adjusted R^2 score.**

***Index Terms*—Community detection, Louvain algorithm, Binary encoding, RNN, LSTM, Cross-validation, R^2 score, MAE**

I. Introduction

Statistical models such as ARIMA, SARIMA, GARCH[29], ARCH[2] etc. have been used as conventional methods in various domains for time series forecasting over a long period of time. Analysis of time series considers the past behavior of data and aids in predicting the future values ahead of time. Prediction using time series data has been successful in vast domains such as Finance, Health sector[16], Employment growth,Weather forecasting[18], etc. According to statistics provided by International Civil Aviation Organisation (ICAO)[1], air travel has become an increasingly popular mode of transport in both developed and developing countries. The scheduled services passengers' traffic raised to 4.5 billion in 2019 leading to a 3.6 percentage increase in passengers growth compared to the previous year. Total global departures witnessed an overall amount of 38.3 millions of passengers in 2019. Forecasting the passenger counts could benefit the passengers to optimally schedule their itinerary as well as the operating airlines to meet the raising demands. Various Airline companies operating all around the world, can plan their operating schedule based on demand through these results and stay ahead in this competitive business. Similarly, travel agencies can show accurate future insights on air travel trends to their customers.

In this paper, we have analyzed over 30 years of data for United States Domestic and International travel and predicted the number of passengers based on different seasons and hubs. We examined the data based on different seasons namely: spring, summer, autumn, winter and implemented time series forecasting respectively for every season. As air travel is seasonal in nature, through this approach, we can leverage the prediction accuracy for the four quarters of any given year. In addition, we wanted to identify the busiest hubs and strongly connected routes in order to provide more precise passenger traffic predictions.

For the purpose of identifying hubs, we carried out the community detection algorithm[24] to group cities into various clusters. Louvain algorithm, one of the popular community detection methods is used to create communities and explore the connectivity among various city routes. In any given network, the nodes (in this case airports) will be strongly connected among their own community and loosely connected between other communities.

A plethora of conventional models exist for time series prediction. However in recent times, LSTM model[28], a form of Recurrent Neural Network(RNN) has been used widely given its nature to hold long term dependencies in sequential data. Therefore, we have implemented the LSTM model for forecast-

[1]https://www.icao.int/Pages/default.aspx

ing the total number of domestic and international passengers. We have evaluated our model's performance using standard error metrics such as Mean Absolute Error(MAE), R^2 score and adjusted R^2 score.Although R^2 score and adjusted R^2 score may seem similar, adjusted R^2 score considers the explanatory variables and adjusts the total number of independent features accordingly.

The remainder of the paper is organized as follows. Section II contains information on the state-of-art methods and algorithms for community detection and LSTM. Section III provides details about the dataset used and the pre-processing techniques adopted in this paper. Section IV explains in detail about Network analysis, Community detection, k-fold cross validation[4] and LSTM model. Section V analyses the technical configuration used in our paper. Section VI discusses about the results obtained at each phase of our experiment.On the other hand Section VII provides an overall insight of our project and elaborates on future works that can be followed by other researchers.

II. Literature Review

A. Network analysis

Song, et al., [24] study the air transport network of 1,060 airports for three airline alliances using to understand the air transport and logistic connectivity between cities and countries. This is achieved using a Social Network Analysis(SNA) methodology as this approach gives more significance to the relationship between nodes in the network rather than the characteristics of individual nodes. The authors used external indicators covering 173 countries to analyse the network rather than utilizing the airport's internal indicators[2]. Wu, W., et al., [27] proposes route-based passenger traffic analysis to identify communities in an airline network as the most significant cities and the routes in the air travel network can be identified with city-wise and route-wise precision. For evaluation purpose, the authors implemented Clauset-Newman-Moore(CNM) algorithm to efficiently calculate the modularity of the communities created out of the airline network. Some of the cities' airports that are found to be key nodes in the community using CNM algorithm include Chicago, Denver and Philadelphia which is relatable as Chicago's O'Hare Airport is the second busiest airport in the world according to Chicago Business Journal's article published in 2021[3]. Lamosa, D., et al.,[20] studied the changes of community structure of Brazil's Sao Jose dos Campos airport on a business day by splitting the travel routes of this airport into 55 traffic zones. There are four large communities identified in the Central, North, South and East sides of the city and also found that more number of passengers fly between 7am and 5pm on any given business day. The authors also suggested that aggregating and observing communities based on space-time dynamics can help in improving urban mobility.

In [13], the authors have made an analysis of the global Air Transportation Network and have examined their significance on the growth of economy. Their results prove that strongly connected cities does not have to be of high centrality. The authors have demonstrated the role of a city on a global scale which depends on its connection patterns both within and outside its community. He, M., et al.,[15] have proposed a methodology to detect communities which are of statistical importance in weighted networks which exhibit a self looping property.The authors have detected communities of three types namely: monads, non-nodal and nodal each representing different regions. Through this approach, US counties that are accumulated and overlapped are identified by demonstrating the commuting patterns among counties of US as a weighted network. The results that were achieved using the CCME-SL approach considering commuting patterns found in the intra-counties were far more efficient that that of the CBSA based methods. Gopalakrishnan, et al., [12] have grouped the air traffic delay state using community detection to identify the characteristics of flight delays. The authors have used mathematical metrics like eigenvector centrality, hub scores and authority scores to analyse the association of nodes within communities. In this paper operational characteristics of the air transportation network have been identified to categorise the type of delay and make changes in their operations respectively.

Chaudhari, et al.,[8] used Random Matrix Theory to analyse the price changes of various cryptocurrencies and implemented community detection to group similar cryptocurrencies with the help of minimum spanning tree. By means of cross-correlation dynamics, they have found that cryptocurrencies within same community exhibit similar characteristics and follow same trend. Lin, et al., [21], in the context of power grids, discussed three types of community detection algorithms namely divisive algorithm, agglomerative algorithm and optimization. Divisive method is a top-down approach i.e., stronger edges(edges with higher weight) are removed first and then, weight of the left-over edges in the network are re-calculated. This method is repeated until strongly inter-connected network is obtained. One such algorithm that follows Divisive approach is Girvan-Newman[11] algorithm. Agglomerative method follows bottom-up approach i.e., only nodes are considered in the network initially. Then, the stronger edges are added first to the nodes following which, the remaining weaker edges are added to build the network completely. Louvain algorithm is one of the algorithms that follows Agglomerative approach for community detection. The difference between these two algorithms is that Divisive algorithm deals with removing the edges, whereas Agglomerative algorithm joins similar nodes of the network into a single community. Optimisation deals with increasing the value of an objective function to calculate the strength of detected communities. Blondel, et al., [6] have proposed Louvain algorithm to identify community structure in a network. This approach works by splitting the network into many sub-units or communities and analyse the nodes that are highly interconnected. The main goal of identifying communities is that the nodes within a group should be strongly connected whereas the nodes between different groups should be loosely connected. Aynaud, et al., [3] compare the stability of three community detection algorithms – WalkTrap, Fast Greedy and Louvain algorithms for static networks. In this paper, stability of the network is analysed by keeping only the nodes that are strongly connected and removing the other nodes from the network. Then, the algorithm is evaluated using cost function

[2]external indicator - airport route, internal indicator - volume of passengers traffic and size of the airport

[3]https://www.bizjournals.com/chicago/news/2021/04/22/ohare-falls-to-worlds-second-busiest-airport-in.html

such as modularity and partition edit distance. It is found that Louvain algorithm produced highest modularity among three algorithms.

B. Feature Encoding and Cross Validation

Among various data pre-processing techniques, feature encoding is highly essential to train a model with non-numeric features as well. Feature Encoding [7] is the process of converting categorical features into machine interpretable vector form. But when the number of features is large, it will result in a high dimensional vector and *one hot* encoding is an example of this scenario. Label encoding can also be used to encode nominal categorical features, but sometimes the Neural Network model can get confused that each label has a different weightage and the results may be inaccurate. Cerda, et.al.,[7] have proposed two methods namely: Poisson Matrix Factorization and Min-Hash Encoding to deal with features of high cardinality. They have mentioned that Min-hash encoding would be suitable if scalability is important and Poisson Factorization would serve the purpose of interpretability. As set inclusions are transformed into disproportioned relations in the Min-Hash encoding approach, similarities in strings are found quickly. In [23], the authors have compared seven feature encoding methods namely: One Hot encoding, Ordinal coding,Sum Encoding, Helmert Encoding, Polynomial Encoding, Backward Differencing, Binary Encoding. Among all the above encoding techniques, they obtained best results when they encoded the features with backward differencing and sum encoding. Subsequently they have passed the pre-processed features to an Artificial Neural Network(ANN) model for classification.

Since Geo-spatial locations play an important role in this experiment,various geocode API's were analysed. Panasyuk, et al., [22] used Google's geocoding API to retrieve the latitude and longitude co-ordinates of Twitter users' location from city and country details. The authors have also experimented with passing additional location parameters to the API in order to identify incorrect geocoded locations. In this way, inaccurate locations can be identified and demographics specific to a particular region or group can be determined precisely and this approach is followed in this paper as well.

Cross validation is an important technique used in time series forecasting with the Machine Learning (ML) models to reduce the error rate and train the model more efficiently.The authors [4] have proposed two methods for train and test data split namely: cross-validation and evaluation using the last part of the series for model selection. The test data is not used while training a ML model and so when the model sees a new data, it might not provide the expected result.Therefore, to make use of all the data available k-fold cross validation is used. In this method, the overall data is divided into 'k' partitions and the model is trained 'k' times on each partition. By this way, the model can handle unseen data better and will also result in a faster performance. Tandon et.al [25] have used 10 fold cross validation to split their data into multiple time series for bitcoin price prediction. They have passed their data into a LSTM model, Random Forest and Linear Regression model and compared the results of each model with the standard Error metric MAE.

C. LSTM for Time Series

Deetchiga et al., [9] explored future trends of air travel time series using Holt-Winters Exponential Smoothing, as the air travel data is seasonal and cyclic in nature. They have predicted yearly passenger count on global data using time series methodology. Apart from conventional time series models,LSTM is being used more for forecasting recently [28]. In Ahmadzade, et al., [1], the authors implemented Augmented Dickey-Fuller(ADF) test for the data of Mehrabad airport, Iran to check whether the time series is stationary. The typical characteristic of a stationary time series is when the mean and variance is constant over time. The authors removed trend and seasonality from data in order to avoid false regression values. It can be seen that coefficient of determination R^2 has increased by 85% for stationary time series data when compared to non-stationary data. Hamami, et.al [14] have built a LSTM model for forecasting the amount of five air pollutants emitted into the atmosphere based on IoT sensors data. The model is also compared with ARIMA to analyse error rates. In this paper, hyperparameter tuning for LSTM is performed and the model is tested with different number of epochs and dropout layers. They achieved best RMSE score of 5.58 when using 20 epochs and 0.2 dropout rate.

Kumar, et. al., [19] have built an LSTM model for predicting the workload in cloud data centers.They have addressed the challenges such as scaling the resources as per the requirement and keeping track of consumption of power. The LSTM model is built in conjunction with a resource manager and the output of the model is fed into it.The resource manager analyses the resource availability at present in the data center before allocating new resources. The performance of the model is evaluated using Mean Squared Error(MSE) and then compared with the back propagation algorithm. Bianchi, et al., [5] suggests that dividing training data into mini-batches and calculating the gradient over the subgroups can efficiently minimise the overall loss function of LSTM. This paper discusses about tuning different hyperparameters such as learning rate, batch size, optimizer, etc. The authors have also provided an overview of how increasing the depth of the neural network model helps in increasing the memory retainability of the network. Liu et.al., [28] implemented multi-step ahead forecasting using LSTM model. The authors have followed two approaches namely iterate-based and direct-based methods. In the first method, output of the previous step is given as one of the inputs in the next step whereas in the second method, the data is passed to multiple models where a model predicts values for single time step, another model predicts for two time steps and so on. For multi-step forecasting, the authors have concluded that LSTM model outperformed ARIMA and General Regression Neural Network (GRNN).

III. Data Description

A. Dataset

This project involves the usage of two datasets – US Domestic and International Travel Statistics respectively. The domestic US travel data was retrieved from the Bureau of Transportation Statistics website[4], which comprises 15 years of data from

[4]https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=GED

2006-2019. The total number of records amount to 3.8 million approximately. The dataset consists of important information such as the period of travel, source and destination, total number of passengers travelled, the airline that carried the passengers and so on. The US passengers' data for International travel is retrieved from U.S. Department of Transportation website[5] and it is available publicly. The dataset contains around 682k records and 16 columns and it covers a period of 30 years from 1990 - 2019. The dataset includes information like number of passengers travelled, source city airport ID, destination city airport ID and date of travel.

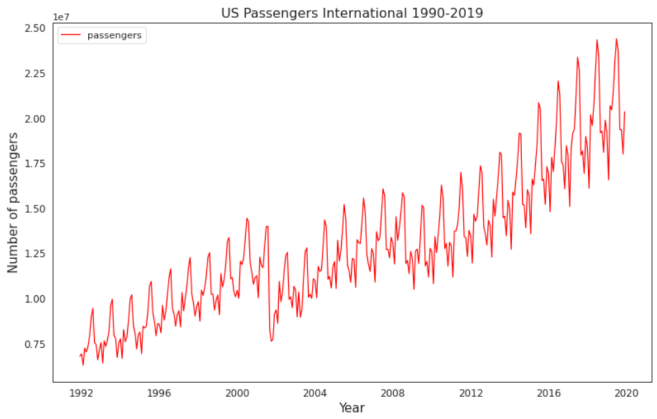

Fig. 1. Raw data - US International Passengers from 2006-2019

B. Pre-processing

With respect to air passenger time series (both Domestic and International US), certain pre-processing methods[26] were implemented for each of the dataset. First, the time series was analysed if it had any trend/ seasonal patterns. Next, a Dickey-Fuller test[1] was applied to the series to check if it was stationary. As it had a unit root, a log transformation was applied to de-trend the series and then, differencing was performed on the logged series to remove the seasonality from data. Figure 1 represents the raw data of International travel from the US.

As part of preparing the domestic US travel dataset before feeding into the model, certain pre-processing techniques were performed. Some columns of lesser importance such as Airport IDs and Carrier names were dropped, as the main focus was on the total passenger count prediction. The latitude and longitude columns were computed with the help of Geopy package[22], for visualisation purpose. Other operations such as changing the data types, renaming and reordering the columns were done.

For the case of the International US travel dataset, as city and country of the airports were missing, they were taken from an additional dataset in Bureau of Transportation Statistics website[6] and mapped to the respective airport IDs. There are around 800 cities to which passengers have flown from USA. Geopy package in Python is used here as well to retrieve the latitude and longitude details of the cities. After handling missing values and data type conversions, there are 682,369 records obtained in our dataset for International US passengers travel.

[5]https://data.transportation.gov/Aviation/International_Report_Passengers/xgub-n9bw

[6]https://www.bts.gov/topics/airlines-and-airports/world-airport-codes

The dataset was then divided into 4 partitions based on the seasons: spring, summer, autumn and winter for seasonal comparisons to witness various patterns in the passenger traffic throughout any given year. Figure 2 demonstrates the process flow of our implementation.

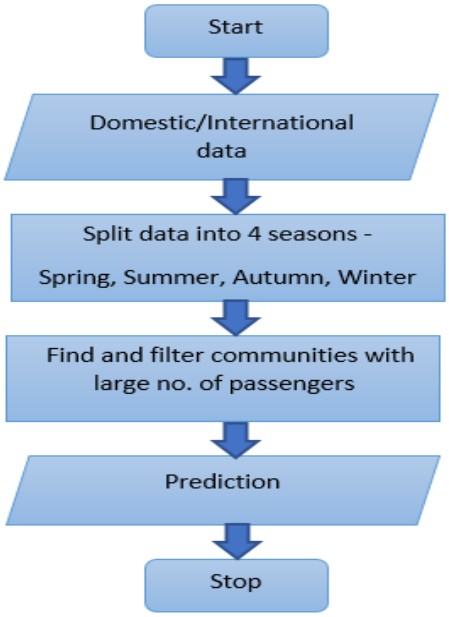

Fig. 2. Experimental Work Flow

IV. Methodology

A. Network Generation

In order to aid community detection in the airport network, the data is converted in the form of graphs. NetworkX package in python was used to create Multi Graphs[21] to represent all edges between any two nodes in the network and provide a graphical form. Cities are represented as nodes, flight routes are represented as edges and the number of passengers are represented as weight of edges. The adjacency matrix of a graph is represented in equation 1:

$$A_{ij} = \begin{cases} w_{ij}, & \text{if } (i,\ j) \in E, \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

From the above equation, it can be seen that a symmetric adjacency matrix is generated for undirected graphs.

B. Community Detection

In order to understand the structure of intricate networks and analyse the connectivity between the node groups, community detection is used. Many community detection approaches[21] such as Divisive, Agglomerative and Optimisation are available. In this paper, we have implemented an Agglomerative approach using the Louvain algorithm as it is suitable for large networks with more number of nodes and weighted edges. This algorithm was developed by Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre [6] in 2008. In this approach, each node is assigned to a separate partition and therefore, the number of partitions will be equal to number

of nodes. Then, each node is placed in different communities and checked until optimum modularity[3] scores are obtained. Next step in Louvain algorithm is to aggregate the nodes present within the same community into one large node. This large node is generated as a sum of all the links present inside the community. This process is iterated until there is no notable increase in modularity (see below). The network exhibits strong inter-connectivity among the nodes within the same community and weaker connectivity between nodes in different communities. Figure 3 explains the working method of Louvain algorithm.

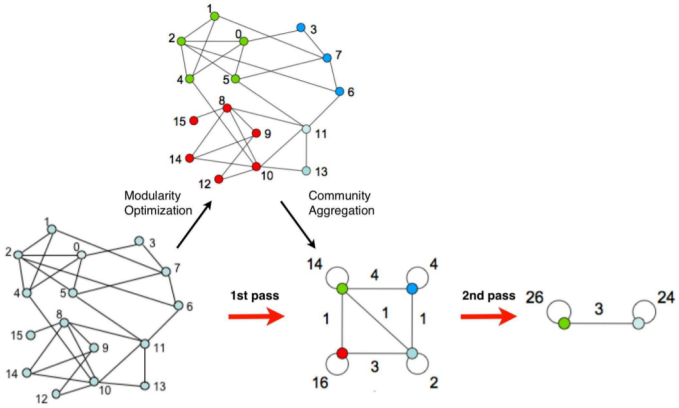

Fig. 3. Working of Louvain algorithm [6]

Modularity is the most common metric used to evaluate how efficiently the network is partitioned. Modularity score falls in the range of -1 to 1 and if the score is higher, we can interpret that the nodes in the communities are strongly connected. The modularity score is calculated using equation 2:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i,\ c_j), \tag{2}$$

where m is the total sum of all edge weights, A_{ij} is the weight of edge from node i to node j, k_i is the total degree of node i summing all the weights of edges that linking node i and c_i denotes the community label to which node i belongs. $\delta(c_i, c_j)$ is a simple delta function.

C. LSTM

Recurrent Neural Networks (RNNs) are renowned for their capabilities to retain information about previously given inputs[5]. Long Short Term Model (LSTM) is a form of Recurrent Neural Networks, which is used to retrieve information from sequential data and remember long term dependencies. It plays a major role in forecasting real time data ranging from IOT systems[14], industrial systems[19] and healthcare[16].

Hochreiter and Schmidhuber[17] were the first to propose the LSTM model in 1997. LSTM differs from the RNNs by their memory cell states, which have the ability to store long or short term information in them. LSTMS are widely used to avoid the vanishing gradient problem. LSTM has three important components: the forget gate, input gate, and the output gate. Figure 5 represents LSTM architecture:

The forget gate decides if the information is important or not and then pushes into the cell state. The input gate generates an

Fig. 4. Community detection for busiest hubs - Boston, Denver and Seattle in Summer; colour of the edges represent flights to different communities

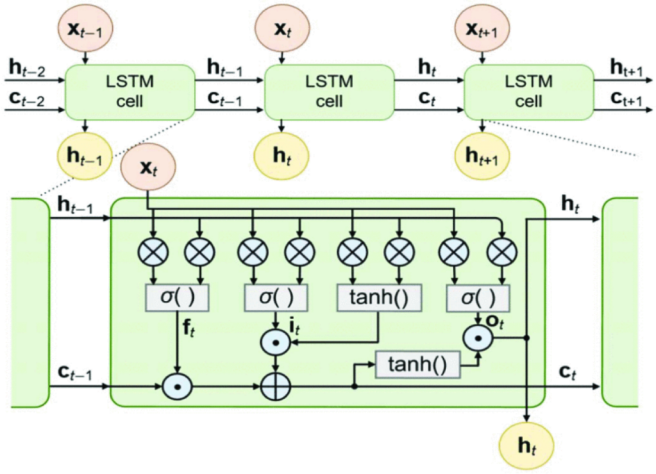

Fig. 5. LSTM Architecture [17]

output of 0 to 1 and chooses the values that are to be updated. Finally the output gate generates the output, which will depend on the cell state. A Sigmoid function is used to decide on the important values of the cell state, which will be provided as output. The following equations are used in the computation of outputs for every gates of the LSTM model:

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{3}$$
$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{4}$$
$$c_t = \tanh(W_c.[h_{t-1}, x_t] + b_c) \tag{5}$$
$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{6}$$
$$h_t = o_t * \tanh(c_t) \tag{7}$$

where x_t: input vector, ht: output vector, c_t: cell state vector, f_t: forget gate vector, it: input gate vector, o_t: output gate vector and W, b are the parameter matrix and vector.

In our dataset, we have two important features: source and destination city which are categorical in nature and cannot be fed directly into the LSTM. Hence, we chose to apply binary encoding[23] to those features due to their high cardinality nature and the resultant feature dimensions could be less. Once the features were encoded, the data was then sorted based on the period of travel.

In order to maintain the features on the same scale, the data is normalized using the MinMax Scaler [4] before passing it to the LSTM model. In general, the dataset will be split into training and test sets in a ratio of 70 to 30. We decided to split our data into train, validation and tests sets so that the model can learn from the validation if it performed well before moving on to the next iteration for training. As our dataset was very large, to account for a faster performance, we wanted to split our dataset into multiple batches. Therefore, k-fold cross-validation [25] was used to divide the data into 'k' partitions and then train the model repeatedly on each partition. Figure 6 shows the working of k- fold cross validation with LSTM:

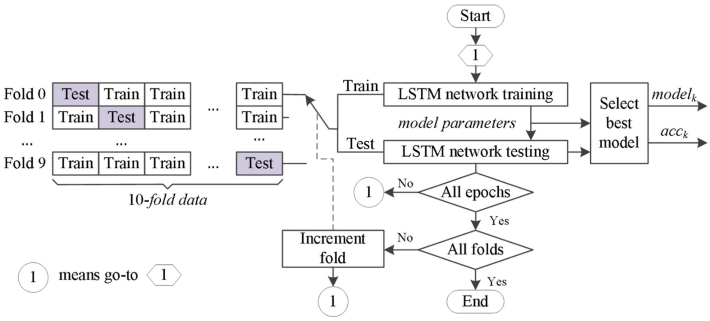

Fig. 6. Cross Validation with LSTM [10]

We experimented on identifying the best number of partitions 'k' and our model provided the best results when k was set to 10. Hyper parameters such as the number of epochs, layers and batch size were tuned for optimization. Standard evaluation metrics such as Mean Absolute Error(MAE), R^2 score and adjusted R^2 score were used to validate the model performance.

V. Experimental Setups

A. Network creation

The latitude and longitude coordinates were essential for visualising the communities in a global map. Hence, we used the Python Geopy and Nominatim [22] packages to retrieve the geo-positional details for each city in our dataset. In order to create a graphical representation of the city pairs present, NetworkX package was used to create the airport nodes and various routes between them. Louvain algorithm was imported from the Community package to create multiple partitions.

B. LSTM

For the purpose of categorical features encoding, category encoders package was installed to apply binary encoding to features such as city pairs. The data was then split into training, validation and test sets in the ratio of 55:25:20 respectively. In order to process the data in mini batches, k-fold cross validation is used. The optimal value of 'k' was determined to be 4 for Domestic US passengers dataset and 10 for International US passengers dataset. After multiple experiments, the final configuration of LSTM hyperparameters[14] is as follows: (i) number of epochs: 50, (ii) learning rate: 0.001, (iii) batch size: 64, (iv) optimiser: Adam, (v) dropout: 0.2, (vi) weight-decay: 0.000001, (vii) number of hidden layers: 64, (viii) input dimension: 21, (ix) output dimension: 1.

C. Evaluation Metric

We have used three standard error metrics[29] in our paper namely: MAE, R2 score and adjusted R2 score. MAE in general is suitable for average performance errors and evaluation based on dimensions. It calculates the average of, the difference between ground truth and model's prediction results and takes absolute of the final value. MAE penalises the outliers present in data. Equation 8 represents the mathematical representation of MAE:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y}| \tag{8}$$

where actual value is represented as y, predicted value is represented as $\overline{y}$.

The R^2 score gives the ratio of variance present in the variables that are dependent. R^2 score has a scale free property. Hence, even for large scaled input values is, the value of R^2 score ranges from 0 to 1. R^2 can be calculated via the following equation 9:

$$R^2 = 1 - \frac{\sum_(y_i - \widehat{y})^2}{\sum_(y_i - \overline{y})^2} \tag{9}$$

where actual value is represented as y, predicted value is represented as $\overline{y}$.

Adjusted R^2 score is similar to R^2 score but with a slight change. In this methods, the total number of independent features are adjusted and it considers the explanatory variables as well. The adjusted R^2 score value will either be lesser or same as the R^2 value and can be computed with the help of the below equation 10.

$$R^2_{adj} = 1 - [\frac{(1-R^2)(n-1)}{n-k-1}] \tag{10}$$

where n is the total number of samples and k is the total number of independent features.

D. Visualisation

In order to visualise the communities, Gephi visualisation tool version 0.9.2 is used with the following plug-ins: Map of Countries and Geo Layout. In addition, Matplotlib package is also used for exploratory data analysis and results visualisation.

VI. Results and Discussion

In this section, results obtained for air passenger traffic forecasting from community detection and LSTM methods are demonstrated. The final data after pre-processing, amounted to 400k records and 3.8M records for International and Domestic US travel respectively. These data are further split based on four different seasons. Once Louvain algorithm is applied for Community detection, 11 communities are created for International network and 9 communities are created for Domestic network. These communities are evaluated using Modularity score and we achieved a score of around 0.3-0.4 for all the

TABLE I
THE COMMUNITIES CREATED FOR FOUR DIFFERENT SEASONS - SPRING, SUMMER, AUTUMN AND WINTER FOR BOTH DOMESTIC AND INTERNATIONAL AIR TRAVEL DATA IS MENTIONED IN THE TABLE.

	Domestic	International
Spring		
Number of communities	9	11
Number of cities in top community	566	244
Number of passengers in top community	4,563,081,721	448,548,116
Modularity	0.3398	0.4001
Summer		
Number of communities	9	11
Number of cities in top community	414	255
Number of passengers in top community	3,209,577,565	513,999,744
Modularity	0.3452	0.3715
Autumn		
Number of communities	9	11
Number of cities in top community	423	347
Number of passengers in top community	3,003,181,796	525,954,439
Modularity	0.3463	0.3298
Winter		
Number of communities	9	11
Number of cities in top community	548	299
Number of passengers in top community	2,784,454,180	391,399,462
Modularity	0.3532	0.4185

seasons, which signifies that the communities are strongly knitted. Table I shows detailed results including number of cities in top community, number of passengers and modularity scores obtained from community detection algorithm. The top communities had around 150 to 200 cities that belongs to a particular geographic region. This implies that certain countries are buzzing with more number of passengers than others. Figure 7 shows the top five communities detected for Domestic US dataset - Summer season.

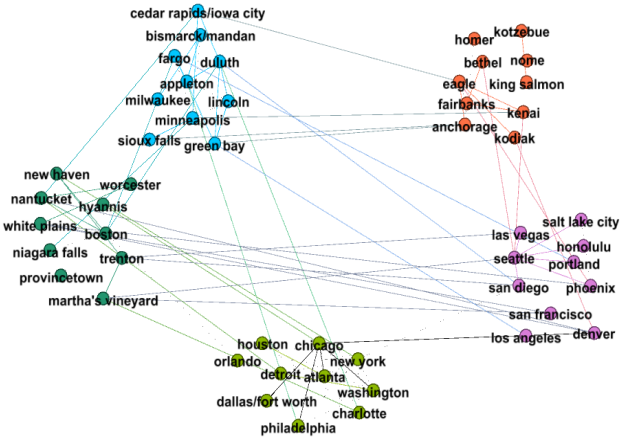

Fig. 7. Community detection for Domestic US passengers - Summer

After close examination, the summer season was found to have the highest number of Domestic and International passengers. This might be due to the vacation period and good weather conditions, as people tend to travel frequently compared to other season.In addition, certain cities have more

TABLE II
ERROR METRICS RESULTS FOR FOUR SEASONS - US DOMESTIC AND INTERNATIONAL.

	MAE	R^2 score	Adjusted R^2 score
Domestic			
Spring	3949.00	0.653	0.651
Summer	4019.95	0.689	0.684
Autumn	2830.72	0.794	0.792
Winter	3608.42	0.681	0.680
International			
Spring	6084.57	0.654	0.653
Summer	5736.54	0.697	0.695
Autumn	5736.54	0.698	0.696
Winter	6009.65	0.699	0.694

number of passengers in specific seasons alone.One such case is Boston city, which has more number of passengers travelled to International destinations in summer.

From the communities obtained, only the top communities which has total number of passengers more than 100 is filtered and then passed on to LSTM for prediction. The model is trained along with k-fold cross-validation in order to separate the time series into multiple mini-batches which helps to decrease the error rate. ”k” value of 4 is used for Domestic dataset whereas ”k” value of 10 is used for International dataset, i.e., the dataset is split into 4 groups and 10 groups and then the model is trained and tested on these sub-groups. Figure 8 shows detailed comparison of actual and the predicted values for all the four seasons for both Domestic and International travel. The performance of the LSTM model is then evaluated using three standard error metrics namely: MAE, R^2 score and adjusted R^2 score. Figure 9 lists the R^2 scores and adjusted R^2 scores we obtained when experimenting with different values of k in cross validation.

VII. CONCLUSION

In this paper, we have proposed a novel approach by using community detection to spot the strongly connected airport routes and then make predictions using LSTM model for busiest hubs, that are identified using Louvain algorithm[6]. Previous works[27][15] [20]involved detecting communities or using machine learning models for forecasting, but they did not make use of community patterns for predictions. Since this methodology gives detailed granular insights, airlines can greatly benefit by planning city-specific and season-specific flight schedules.Our analysis showcased large number of passengers travelled especially in summer, as it is a vacation period and carrier companies can schedule more flights for busiest cities and routes in this time, which can help to increase their revenue.In particular, Boston,Philadelphia and Detroit were found to be the busiest cities during summer.In this way, hotspots of the air travel network for spring, autumn and winters seasons are also identified and then the future number of passengers were predicted.

With regard to LSTM, we used 4 fold and 10 fold cross validation for Domestic and International US dataset respectively.LSTM with cross validation performed efficiently and provided a R^2 score of 0.794, whereas vanilla LSTM provided R^2 score of 0.2 only.In future, more algorithms like Girvan-Newman algorithm can be extended for community detection

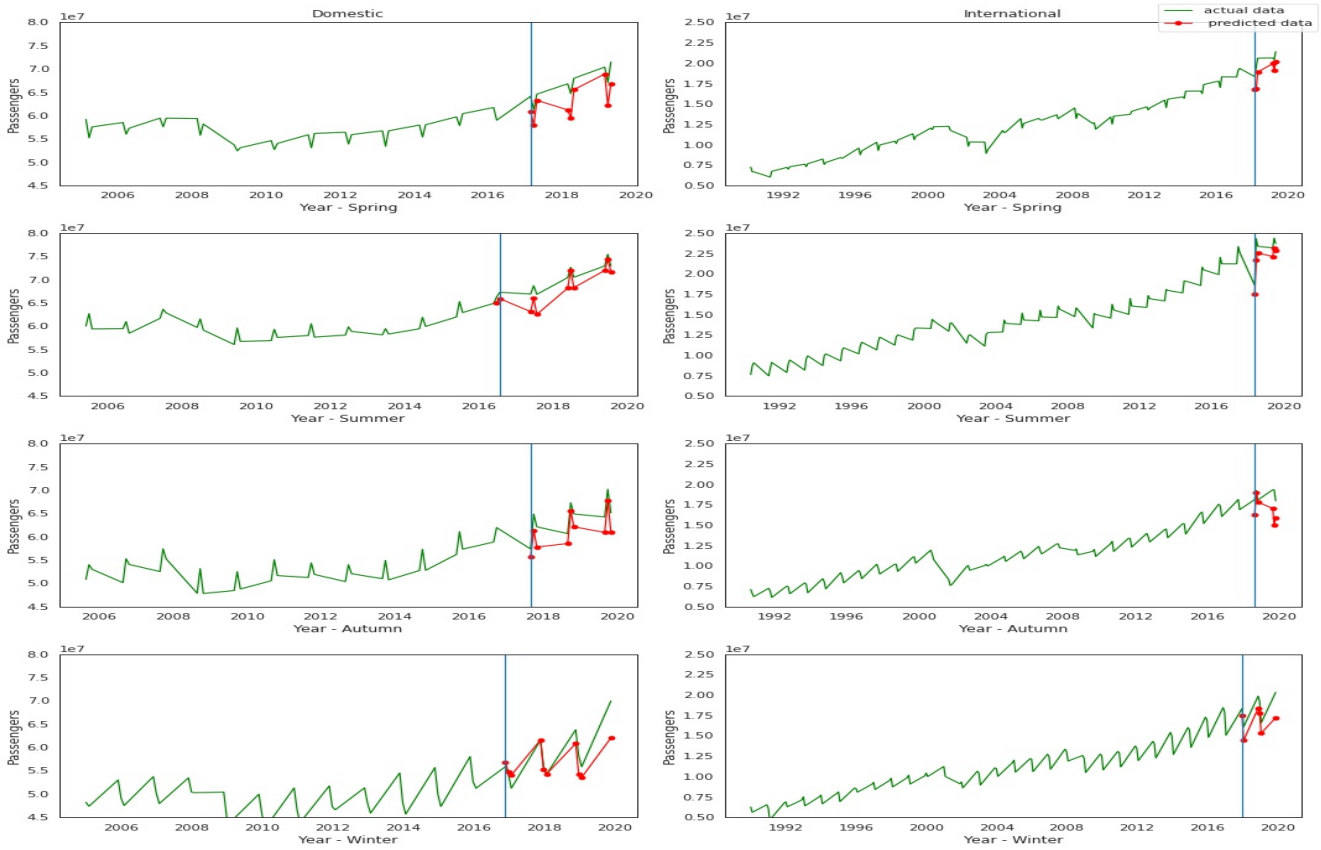

Fig. 8. Actual vs Predicted number of passengers - Domestic and International

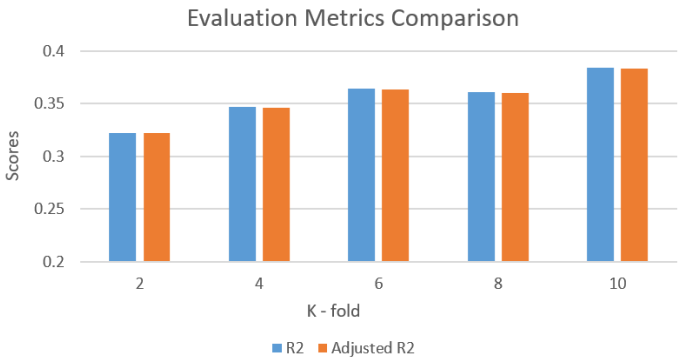

Fig. 9. International Spring US - Evaluation metrics comparison for different k-folds

to create strongly interconnected network with increased modularity and LSTM with multi-steps can be implemented for prediction.

REFERENCES

[1] Ahmadzade, F. (2010). Model for forecasting passenger of airport.

[2] Alam, M. Z., Siddikee, M., and Masukujjaman, M. (2013). Forecasting volatility of stock indices with arch model. *International Journal of Financial Research*, 4.

[3] Aynaud, T. and Guillaume, J.-L. (2010). Static community detection algorithms for evolving networks. In *8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*, pages 513–519.

[4] Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213. Data Mining for Software Trustworthiness.

[5] Bianchi, F. M., Maiorino, E., Kampffmeyer, M. C., Rizzi, A., and Jenssen, R. (2017). An overview and comparative analysis of recurrent neural networks for short term load forecasting. *CoRR*, abs/1705.04378.

[6] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

[7] Cerda, P. and Varoquaux, G. (2020). Encoding high-cardinality string categorical variables. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

[8] Chaudhari, H. and Crane, M. (2020). Cross-correlation dynamics and community structures of cryptocurrencies. *Journal of Computational Science*, 44:101130.

[9] Deetchiga., S., harini, U. K., Marimuthu, M., and Radhika, J. (2018). Prediction of passenger traffic for global airport using holt's winter method in time series analysis. In *2018 International Conference on Intelligent Computing and Communication for Smart World (I2C2SW)*, pages 165–169.

[10] Faust, O., Barika, R., Shenfield, A., Ciaccio, E. J., and Acharya, U. R. (2021). Accurate detection of sleep apnea with long short-term memory network based on rr interval

signals. *Knowledge-Based Systems*, 212:106591.

[11] Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. 99(12):7821–7826.

[12] Gopalakrishnan, K., Balakrishnan, H., and Jordan, R. (2016). Clusters and communities in air traffic delay networks. In *2016 American Control Conference (ACC)*, pages 3782–3788.

[13] Guimerà, R., Mossa, S., Turtschi, A., and Amaral, L. A. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. 102(22):7794–7799.

[14] Hamami, F. and Dahlan, I. A. (2020). Univariate time series data forecasting of air pollution using lstm neural network. In *2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, pages 1–5.

[15] He, M., Glasser, J., Pritchard, N., Bhamidi, S., and Kaza, N. (2020). Demarcating geographic regions using community detection in commuting networks with significant self-loops. *PLOS ONE*, 15:e0230941.

[16] Islam, M., Umran, H., Umran, S., and Karim, M. (2019). Intelligent healthcare platform: Cardiovascular disease risk factors prediction using attention module based lstm. pages 167–175.

[17] Jo, J., Kung, J., Lee, S., and Lee, Y. (2019). Similarity-based lstm architecture for energy-efficient edge-level speech recognition. In *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, pages 1–6.

[18] Kothapalli, S. and Totad, S. G. (2017). A real-time weather forecasting and analysis. In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, pages 1567–1570.

[19] Kumar, J., Goomer, R., and Singh, A. K. (2018). Long short term memory recurrent neural network (lstm-rnn) based workload forecasting model for cloud datacenters. *Procedia Computer Science*, 125:676–682. The 6th International Conference on Smart Computing and Communications.

[20] Lamosa, J. D., Tomás, L., Quiles, M., Londe, L., Santos, L., and Macau, E. (2021). Topological indexes and community structure for urban mobility networks: Variations in a business day. *PLOS ONE*, 16:e0248126.

[21] Lin, G., Liu, S., Zhou, A., Dai, J., Chai, B., Zhang, B., Qiu, H., Gao, K., Song, Y., and Chen, R. (2017). Community detection in power grids based on louvain heuristic algorithm. In *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*, pages 1–4.

[22] Panasyuk, A., Yu, E. S.-L., and Mehrotra, K. G. (2019). Improving geocoding for city-level locations. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 416–421.

[23] Potdar, K., Pardawala, T., and Pai, C. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, 175:7–9.

[24] Song, M. G. and Yeo, G. T. (2017). Analysis of the air transport network characteristics of major airports. *The Asian Journal of Shipping and Logistics*, 33(3):117–125.

[25] Tandon, S., Tripathi, S., Saraswat, P., and Dabas, C. (2019). Bitcoin price forecasting using lstm and 10-fold cross validation. In *2019 International Conference on Signal Processing and Communication (ICSC)*, pages 323–328.

[26] Virili, F. and Freisleben, B. (2000). Nonstationarity and data preprocessing for neural network predictions of an economic time series. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 5, pages 129–134 vol.5.

[27] Wu, W., Haoyu, Z., Zhang, S., and Witlox, F. (2019). Community detection in airline networks: An empirical analysis of american vs. southwest airlines. *Journal of Advanced Transportation*, 2019:1–11.

[28] Yunpeng, L., Di, H., Junpeng, B., and Yong, Q. (2017). Multi-step ahead time series forecasting for different data patterns based on lstm recurrent neural network. In *2017 14th Web Information Systems and Applications Conference (WISA)*, pages 305–310.

[29] Zhang, W. and Wang, Y. (2011). Garch family model based on the shanghai stock market shorting mechanism analysis. In *2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, pages 5046–5049.