Flight Delay Prediction using Binary Classification

Roshni Musaddi [1], Anny Jaiswal [2], Pooja J [3], Mansvi Girdonia [4], Minu M.S [5]
[1, 2, 3, 4] Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.
[5] Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India.

Abstract – Nowadays, aircrafts have become a necessity because they easy life. They are efficient in carrying goods and passengers around the world. It also supplies emergencies in warfare and takes a vital role in carrying medical necessities. Hence, advent of airplanes is considered important. Delays in aircrafts can affect thousands of people across the globe either directly or indirectly. There are a lot of reasons of delays in aircrafts such as critical weather, security issues, traffic and many more. There are several methods proposed to predict the flight delays but due to various complexities of the ATFM and the huge datasets involved, it has become very difficult to find an accurate solution for this complication. Many algorithms have been implemented to forecast flight delays. We are using Python in Visual Studio Code. We implement Binary Classification to prepare a model that can predict the delays.

Index Terms – Binary Classification, Visual Studio Code, Regularities, Python, Transportation, Complexities, Air Traffic Flow Management.

1. INTRODUCTION

In the present scenario, there are approximate 2,500 thousand people who travel frequently. Air transportation is considered as the best transportation due to its affordable prices, lesser travelling time, safer than other transports, in-flight entertainment, etc. A delay not only affects the passengers travelling but also affects the medical patients as well as the business organizations, transporting their goods. Whenever a connecting flight delays ,it not only affects its current destination ,but affects the next successive destinations as well. Machine Learning is an AI application, which can improve and learn from its previous experiences. Many machine learning methods have been suggested to anticipate flight delays. This paper aims at developing a Binary Classifier model to forecast delays in airlines accurately. A binary classifier classifies the elements into two groups as true or false. We predict whether a flight will be delayed or not. The datasets have been stored as csv files. One of the Excel files contains information of all the flights of year 2015. The datasets contains more than 5,800,000 flight details. Another csv file contains the description of all the airports. To reduce the time and memory complexity, we are taking into account a portion of the dataset, thereby keeping flights details of January 2015. In the first step, we are cleansing the data sets i.e. cleaning dates and time and removing the unnecessary data. We are basically comparing airlines by the statistical description of other airlines and making a delay distribution by establishing the ranking of airlines. We are implementing the algorithm on Visual Studio Code in python. We are fetching the required data and refining the dataset by removing the unnecessary data. This modified data set is converted into sparse matrix. We are using Grid search on Random Forest model to get the ROC Curve .The results is stored into two groups , 0 when it indicates the flight is on time and 1 when the flight gets delayed. In our paper, we have used various histograms to compare various airlines with respect to days and week and to know which of these airlines serves best in terms of less delays. Our paper, thus by concluding the flight delays gives various options to the passengers before they travel through these airlines.

2. RELATED WORK

Fei Rong[1] designed a factor model of irregular flights by series analysis of factual data, which is united with Bayesian Network (BN) and Gaussian mixture model -expectation maximum algorithm (GMM-EM) algorithm. They tested the data set provided by Air Traffic Management Bureau. He succeeded in predicting the downstream delay with good accuracy, when the delay happened in upstream.

Young Jin Kim[2] used Deep Recurrent Neural Network to discuss architectures of RNN and LSTM network. Also, he showed the benefits of stacking these networks and the ways to make an architecture deeper using RNN. According to him, deep architecture has improved the accuracy of the airport delay prediction models. He showed how a single day delay status can be acquired by applying the deep LSTM RNN architecture.

Kai-Quan Cai[3] investigated a multi-objective air traffic network flow optimization (MATNFO) problem .They developed RTA which is a systematic approach to search the optimal flight routes and departure arrival time. This algorithm searches flight routes and time-slots sequentially and efficiently, and thus avoids high computational overhead. RTA has high computational efficiency which can satisfy the requirement of ATFM. Their results showed that RTA outperforms related work for the MATNFO problem. They

improved various formulation for the air flow traffic management.

Balasubramanian Thiagarajan[4] made a model consisting of two phases - (i) departure delay prediction and (ii) arrival delay prediction to efficiently predict the departure and arrival delays of flights using flight schedule and weather features. At first the model implements binary classification to forecast the occurrence of delays and then performs regression to get the value of the delay in minutes. He concluded that Gradient Boosting Classifier executed the best in classification stage and Extra-Trees Regressor executed the best in regression stage. Also, the departure delay prediction had comparatively higher error rates. Henceforth, they developed a decision support tool (DST) which can serve the dual purpose of helping users in arriving on time for their flights / helping airlines accurately predict when their flights would arrive at the gate.

Hugo Alonso[5] developed a so called unimodal model, to predict the time delays. He implemented the unimodal model with neural networks and implemented binomial model using trees, for comparison purposes. He realized that arrival delay and ground operation time are the most significant variables for departure delay prediction.

Suvojit Manna[6] used the Gradient Boosting model for predicting the delay in a flight. This model has achieved the highest Coefficient of Determination of 92.3185% for the given data in case of arrival and 94.8523% in case of departure. It can be used to predict the delay in flights in various airports of United States of America accurately. Also, this model could be used by people and airline agencies to predict delay in flight accurately. They limited their model with dataset of 70 airports, so it can predict the delays for those airports itself.

Varsha Venkatesh[7] used Neural Network and Deep belief Network for predicting the delay in flights. She used one hidden layer having 3 neurons and got the prediction accuracy of 92% with Neural Network and an accuracy of 77% for Deep belief Network with 4 neurons each in two hidden layers.

3. PORPOSED MODELLING

A. Dataset

The U.S. Department of Transportation's Bureau of Transportation Statistics (BTS) holds the record of all the domestic flights and their on-time, cancelled, delayed and diverted flight performances. The dataset has been collected from their Air Travel Report and contains 2015 flight summary. The dataset includes three csv files:

Flights.csv - It contains 31 columns describing each flight's day, month, year, flight number, airline, scheduled arrival, origin airport, destination airport, scheduled departure, departure time, taxi out, taxi in, scheduled time, departure delay, elapsed time, weather delay, distance, wheels off, tail number, arrival time, arrival delay, diverted, cancelled, cancellation reason, air system delay, security delay, airline delay, late aircraft delay, air time.

Airlines.csv - It contains 2 columns of airlines names and their codes.

IATA_CODE	AIRLINE
UA	United Air Lines Inc.
AA	American Airlines Inc.
US	US Airways Inc.
F9	Frontier Airlines Inc.
B6	JetBlue Airways
OO	Skywest Airlines Inc.
AS	Alaska Airlines Inc.
NK	Spirit Air Lines
WN	Southwest Airlines Co.
DL	Delta Air Lines Inc.
EV	Atlantic Southeast Airlines
HA	Hawaiian Airlines Inc.
MQ	American Eagle Airlines Inc.
VX	Virgin America

Fig. 1: Dataset for the airlines.csv file

Airports.csv - It has 7 columns containing airport's code and name, city, state, country, latitude and longitude. It contains 323 airport's details with their codes.

IATA_CODE	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
ABE	Lehigh Val	Allentown	PA	USA	40.65236	-75.4404
ABI	Abilene Re	Abilene	TX	USA	32.41132	-99.6819
ABQ	Albuquerq	Albuquerq	NM	USA	35.04022	-106.60919
ABR	Aberdeen	Aberdeen	SD	USA	45.44906	-98.42183
ABY	Southwest	Albany	GA	USA	31.53552	-84.19447
ACK	Nantucket	Nantucket	MA	USA	41.25305	-70.06018
ACT	Waco Reg	Waco	TX	USA	31.61129	-97.23052
ACV	Arcata Air	Arcata/Eu	CA	USA	40.97812	-124.10862
ACY	Atlantic Ci	Atlantic Ci	NJ	USA	39.45758	-74.57717
ADK	Adak Airpo	Adak	AK	USA	51.87796	-176.64603
ADQ	Kodiak Air	Kodiak	AK	USA	57.74997	-152.49386
AEX	Alexandria	Alexandria	LA	USA	31.32737	-92.54856
AGS	Augusta Re	Augusta	GA	USA	33.36996	-81.9645
AKN	King Salmo	King Salmo	AK	USA	58.6768	-156.64922
ALB	Albany Int	Albany	NY	USA	42.74812	-73.80298
ALO	Waterloo	Waterloo	IA	USA	42.55708	-92.40034
AMA	Rick Husba	Amarillo	TX	USA	35.21937	-101.70593
ANC	Ted Steven	Anchorage	AK	USA	61.17432	-149.99619
APN	Alpena Co	Alpena	MI	USA	45.07807	-83.56029
ASE	Aspen-Pitk	Aspen	CO	USA	39.22316	-106.86885
ATL	Hartsfield-	Atlanta	GA	USA	33.64044	-84.42694
ATW	Appleton I	Appleton	WI	USA	44.25741	-88.51948
AUS	Austin-Ber	Austin	TX	USA	30.19453	-97.66987

Fig. 2: Dataset for airports.csv file

B. Cleaning

Each row in flights.csv describes a flight and we have more than 5,800,000 flights in the year 2015. The main aspects covered using python are:

- Visualization: matplotlib, seaborn
- Data manipulation: pandas, numpy
- Modeling: sklearn, scipy

The number of flights that were not late or cancelled was much higher than those which were late. Hence, we remove the ones that were not late. Flights from January 2015 were only taken to reduce the size of the dataset. Since days of a year would give us 365 columns, we load it to calculate the week of the year thereby, giving us 52 columns and the 'month' column is removed as it becomes redundant after adding the column for week. The dataset has airport codes but some of these are listed using 5 digits instead of 3 letter codes. These issues were settled later. The departure time has been sectioned into 4 sections of 6 hours each starting at midnight. Arrival-delay, departure-delay and cancelled flights were deleted as they are not required. We also removed data from airports which had much lower number of flights to simplify the statistics.

4. RESULTS AND DISCUSSIONS

Whether a flight is late for more than 60 minutes or cancelled is combined as one feature to avoid complexities.

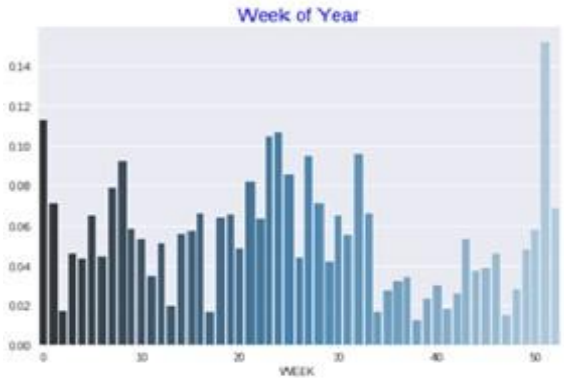

Fig. 3: Plotting week against mean delay

The calculated mean delay is plotted against day of week, week of year, airlines and scheduled departure time.

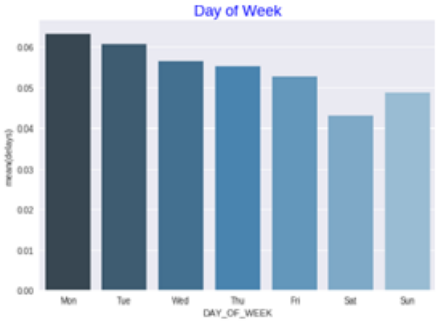

Fig. 4: Plotting days against mean delay

The various airlines are compared to find out which one gets delayed more so that a passenger can have an option to choose a flight which has less chance of getting delayed.

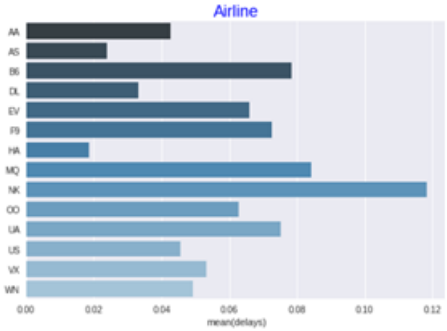

Fig. 5: Plotting mean delay against various airlines

The categorical features were converted into sparse matrices using Label Binarizer which transforms multi-class labels as binary labels. Then we divide the dataset into test and training set. Random Forest Search is a Supervised algorithm which creates many decision trees and finally merges them for more accurate predictions.

Fig. 6: Plotting departure time against mean delay

Grid search was then applied on Random Forest model. Pickle was imported and used to save the outcome of our grid search. Grid search was performed on both training and test set. The true and false predictions (including incorrect predictions) were plotted on the dataset and the results obtained are as follows:

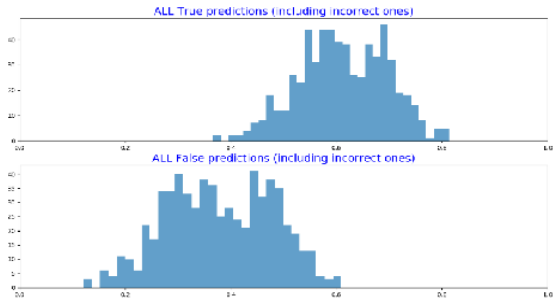

Fig. 7: Histogram for all true and false predictions

The histogram of the results obtained for the training set is as follows:

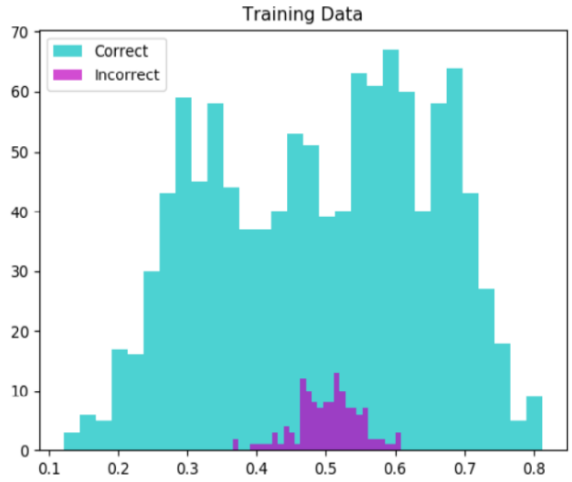

Fig. 8: Histogram for all appropriate and inappropriate predictions for the training data

The histogram of the results acquired for the test set is plotted where the blue color shows the correct predictions and the purple color shows the false predictions.

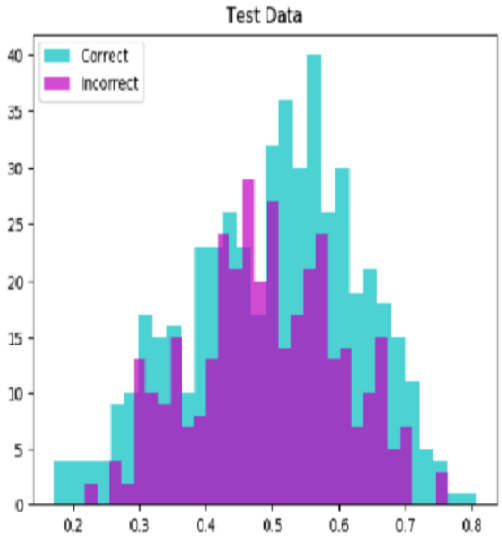

Fig. 9: Histogram for all correct and incorrect predictions of the test data

The ROC curve is a visual representation of the capability of the binary classifier with a certain threshold value. This curve for positive rate against negative rate of the modified dataset was plotted with a threshold value. The positive rate is called sensitivity and the negative rate is called fall-out.

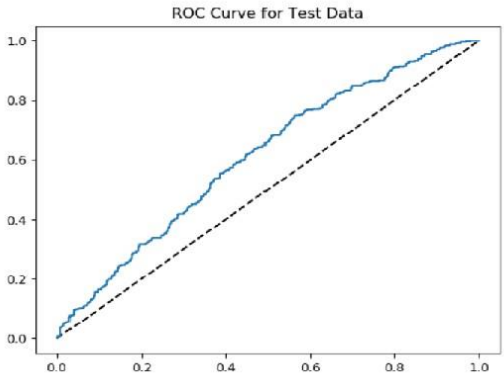

Fig. 10: ROC curve for test data

The dataset was modified to consider only a part of the actual data. Now we plot the same histogram and ROC curve for the whole dataset. The comparison to show the positive predictions on full data was done and plotted to acquire a histogram as follows:

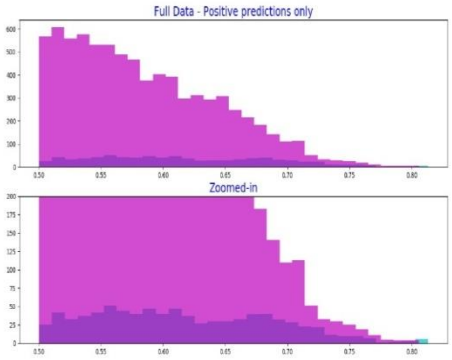

Fig. 11: Histogram for all positive predictions on the whole dataset

The histogram shows that the number of flights being late is very less compared to the number of flights being late or cancelled. The ROC curve for the full data was obtained.

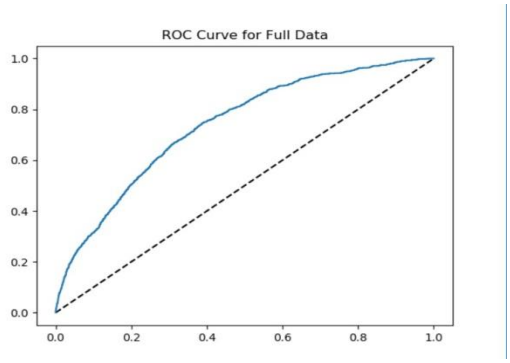

Fig. 12: ROC curve acquired for the whole dataset

This can inform passengers in advance about the chance of a flight being delayed or cancelled compared to other flights. The proposed work aims at analyzing delays in flights using predictions with score higher than 0.85. The ROC curve determines the ability of our Binary Classifier. The obtained curve is quite good though these predictions can be really difficult.

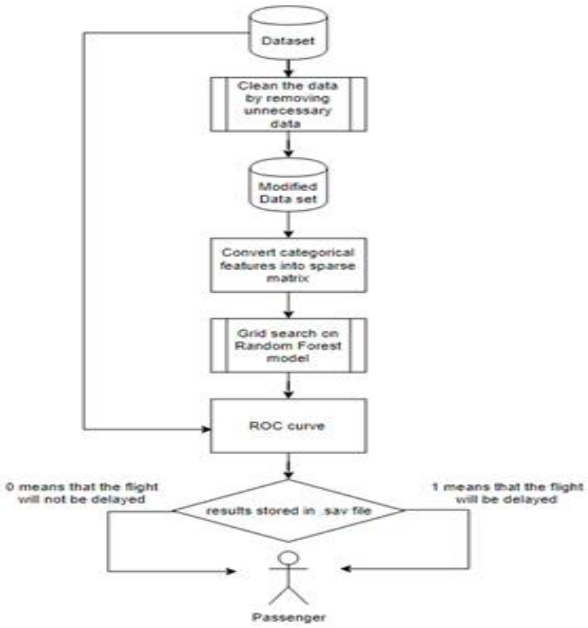

Fig. 13: Flowchart of the proposed model .

5. CONCLUSION

This paper proposes Binary Classification for predicting delay in flight take offs. The datasets from the month of January, 2015 from the Air Travel Report for monthly Consumers of the U.S .Department of Transportations BTS has been taken and filtered to remove unnecessary data like the flights which are not delayed. We have taken datasets of different airlines from different airports to check which one gets delayed more often. The delay can be due to climate conditions or traffic in airspace and airports or any other reason. This gives the passengers the freedom of choosing their best airlines to travel at the extreme conditions. It also excludes error in calculating delays and provides accurate results.

REFERENCES

[1] FeiRong, LiQianya, Hu Bo, Zhang Jing , Yang Dongdong.”The Prediction of Flight Delays based The Analysis of Random Flight Points” .*Proceedings of 34th Chinese Control Conference.July 28-30 ,2015.*

[2] Young Jin Kim, Sun Choi, Simon Briceno and Dimitri Mavris.”A Deep Learning Approach to Flight Delay Prediction”.*Aerospace Systems Design Laboratory .Georgia Institute of Technology. 978-1-5090-2523-7/16/$31.00 ©2016 IEEE*

[3] Kai-Quan Cai, Jun Zhang, Ming-Ming Xiao, Ke Tang, and Wen-Bo Du.”Simultaneous Optimization of Airspace Congestion and Flight Delay in Air Traffic Network Flow Management”.*1524-9050 © 2017 IEEE.http://www.ieee.org/publications_standards/publications/rights/index.html*

[4] Balasubramanian Thiagarajan, Lakshminarasimhan Srinivasan, Aditya Vikram Sharma, Dinesh Sreekanthan, Vineeth Vijayaraghavan[4] . “A Machine Learning Approach for Prediction of On-time Performance of Flights “.*978-1-5386-0365-9/17/$31.00 ©2017 IEEE*

[5] Hugo Alonso and Antonio Loureiro .”Predicting Flight Departure Delay at Porto Airport: A Preliminary Study”.

[6] Suvojit Manna, Sanket Biswas, Riyanka Kundu Somnath Rakshit, Priti Gupta and Subhas Barman .“A Statistical Approach to Predict Flight Delay Using Gradient Boosted Decision Tree”.*2017 International Conference on Computational Intelligence in Data Science(ICCIDS). 978-1-5090-5595-1/17/$31.00 ©2017 IEEE*

[7] Varsha Venkatesh,Arti Arya,Pooja Agarwal,Lakshmi S,Sanjay Balana.”Iterative Machine and Deep Learning Approach for Aviation Delay Prediction*”. 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON). 978-1-5386-3004-4/17/$31.00 ©2017 IEEE*

[8] Cristianini,N. and Shawe-Taylor, J. (2000).Cambridge University Press, United Kingdom, 1st edition.An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.

[9] El-Rabbany, A. (2006). Introduction to GPS: The Global Positioning System. Artech House, Norwood, 2nd edition.

[10] Fernandez-Navarro, F., Riccardi, A., and Carloni, S. ´ (2015). Ordinal regression by a generalized forcebased model. IEEE Transactions on Cybernetics, 45(4):844–857.

[11] Frank, E. and Hall, M. (2001). A simple approach to ordinal classification. In Proceedings of the 12th European Conference on Machine Learning (ECML 2001), volume 1, pages 145–156.

[12] Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York, 2nd edition.

[13] N. Xu, G. Donohue, K. B. Laskey, and C.-H. Chen, “Estimation of delay propagation in the national aviation system using bayesian networks,” in 6th USA/Europe Air Traffic Management Research and Development Seminar. Citeseer, 2005.

[14] J. J. Rebollo and H. Balakrishnan, “Characterization and prediction of air traffic delays,” Transportation Research Part C: Emerging Technologies, vol. 44, pp. 231–241, 2014.

[15] S. Choi, Y. J. Kim, S. Briceno, and D. N. Mavris, “Prediction of weatherinduced airline delays based on machine learning algorithms,” in Digital Avionics Systems Conference (DASC), 2016 IEEE/AIAA 35th. IEEE, 2016.

[16] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, “Deep learning applications and challenges in big data analytics,” Journal of Big Data, vol. 2, no. 1, pp. 1–21, 2015.

[17] H. Kashyap, H. A. Ahmed, N. Hoque, S. Roy, and D. K. Bhattacharyya, “Big data analytics in bioinformatics: A machine learning perspective,” arXiv preprint arXiv:1506.05101, 2015.

[18] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, “Traffic flow prediction with big data: A deep learning approach,” Intelligent Transportation Systems, IEEE Transactions on, vol. 16, no. 2, pp. 865–873, 2015.

[19] K.-i. Funahashi and Y. Nakamura, “Approximation of dynamical systems by continuous time recurrent neural networks,” Neural networks, vol. 6, no. 6, pp. 801–806, 1993.

[20] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013, pp. 6645–6649.

[21] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, “How to construct deep recurrent neural networks,” arXiv preprint arXiv:1312.6026, 2013.

[22] “Bureau of transports statistics, http://www.transtats.bts.gov.”