ISSN 2395-1621

DEPARTURE CAPACITY PREDICTION FOR HUB AIRPORT IN THUNDERSTORM BASED ON DATA MINING METHOD

[#1]Rohan Laxman Joshi, [#2]Dnyaneshwar Ramchandra Jawane,
[#3]Akshay vithal Ithape, [#4]Uday vasantrao Jagtap

[1]joshirohan1997@gmail.com
[2]jawanednyaneshwar1@gmail.com
[3]akshayithape1111@gmail.com
[4]udayjagtap10@gmail.com

[#1234]Department of Computer Engineering

Sinhagad Academy Of Engineering Kondhawa, Pune 411048

ABSTRACT

The primary goal of model proposed in this seminar is to predict airline delays caused by inclement of weather conditions using data mining and supervised machine learning algorithm (Random Forest). 2008 US domestic flight data and weather data was extracted for training and prediction. Four different models were developed to for analysis of behavior of different parameters. Departure and Arrival delays were separately determined. OOB score was calculated to determine optimum number of trees. Sampling techniques (SMOTE) were then applied on data to improve the performance of model. Every model's performance was compared using precision, recall, F1 score, Accuracy, Confusion matrix, and AUC under ROC.

Keywords: Data Science, Data mining, Machine Learning, Delay Prediction, Weather, Imbalanced Training data, Sampling Techniques

ARTICLE INFO

Article History

Received: 10th December 2017

Received in revised form :

10th December 2017

Accepted: 13th December 2017

Published online :

14th December 2017

I. INTRODUCTION

Pothole With the reduction of operating costs in the aviation industry due to fuel efficient aircraft, the cost reduction by leveraging modern technology and the increase in household disposable income, the volume of air travel increased from 450 billion passenger-miles in 1997 to 600 billion passenger-miles in 2014. (BTS, US Passenger Miles Table) [1]

The higher passenger traffic and the increase in the number of flights offered by airlines, means that during bad weather conditions the National Airspace System (NAS) capacity in the United States is challenged to handle the number of scheduled flights. Passengers can book flights up to one year before the departure date, which is usually when an airline publishes its flight schedule. However, the planned and published flight schedule does not account for the potential impact of weather that may occur on the day of the flight. Instead, the schedule is mainly set according to profit and market share considerations. As a result, the imbalance between flight demand and NAS capacity in the US yields flight delays. The average load factor in a flight for domestic operations in 2012 was close to 83%. Flight delays not only cause time loss for passengers but also create multiplicative inefficiency, wreaking havoc downstream by disrupting airport runway operations and the planning of airlines. the annual cost of domestic flight delays to the US economy was estimated to be $31-40billion in 2007 (Joint Economic Committee, US Senate 2008). Correctly predicting flight delays allows passengers to be prepared for the disruption of their journey and allows airlines to pro-actively respond to the potential causes of the flight delay to mitigate their impact. The abundant research efforts from data scientists, researchers, companies and government agencies on airline flight delays confirms that this is an important area. In particular, the main benefits of better flight prediction are significant operational cost savings and a non-negligible improvement in quality of life for those who use air as an important mean of transport. An

accurate online flight delay predictor would certainly generate a lot of interest in the world of air travel.The goal of work done in this seminar is to use exploratory analysis and to develop machine learning models to predict airline's departure and arrival delays. Based on the literature reviews, this type of problem is actively examined by many researchers and GE even brought out a flight quest challenge with an award of $250,000 to the team who can most accurately predict flight delays.

II. LITERATURE SURVEY

The increase in delays in the National Airspace System (NAS) has been the subject of studies in recent years. The literature on delay analysis and its potential remedies extends back over several decades. Levine (1969) argues that pricing is a better means of allocating scarce airport capacity to meet the demand than other mechanisms being considered at the time, such as slot allocation. The Federal Aviation Administration (FAA) describes the increase in delays and 10 cancellations from 1995 through 1999.

Schaefer and Miller (2001) found that the current system for collecting causal data does not provide the appropriate data for developing strong conclusions for delay causes and recommend changes to the current data collection system.

Allan et al. (2001) examined delays at New York City Airports from September 1998 through August 2000 to determine the major causes of delay that occurred during the first year of an Integrated Terminal Weather System (ITWS) use and delays that occurred with ITWS in operation that were "avoidable" if enhanced weather detection. The methodology used in the study has considered major causes of delays (convective weather inside and well outside the terminal area, and high winds) that have generally been ignored in previous studies of capacity constrained airports such as Newark International Airport (EWR). The research found that the usual paradigm of assessing delays only in terms of Instrument Meteorological Conditions (IMC) and Visual Meteorological Conditions (VMC) and the associated airport capacities is far too simplistic as a tool for determining which air traffic management investments best reduces the "avoidable" delays.

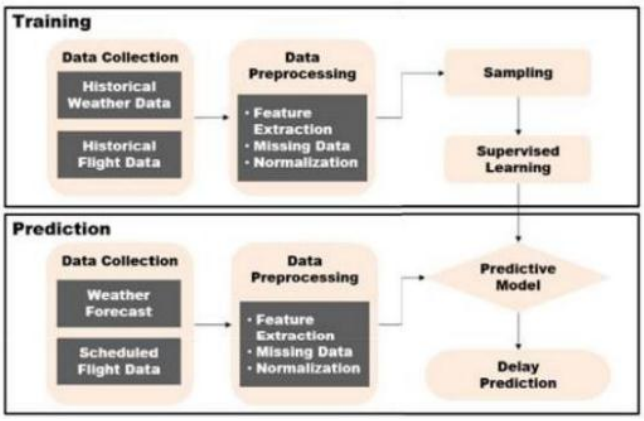
Schaefer and Miller (2001) use the Detailed Policy Assessment Tool (DPAT) to model the propagation of delay throughout a system of airports and sectors. To estimate delays, throughputs, and air traffic congestion in a typical scenario of current operations in the U. S., DPAT models the flow of approximately 50,000 flights per day throughout the airports and airspace of the U. S. National Airspace System (NAS) and can simulate flights to analyze delays at airports around the world. They obtained results for local flight departure and arrival delays due to IMC, propagation for IMC, comparisons to 11 VMC results, and a comparison of propagated delays to entire system.

Rosen (2002) measures the change in flight times resulting from infrastructure-constant changes in passenger demand. Results indicate that delays rise with the ratio of demand to fixed airport infrastructure, decreasing average flight times by close to seven minutes after the sharp decrease in demand in the Fall of 2001. Flight time differences between the airlines in the sample are small, though the larger United had shorter average flight times in the winter quarter than America West, the smaller airline in the data sample.

Janic(2003) presents a model for assessment of the economic consequences of large-scale disruptions of an airline single hub-and-spoke network expressed by the costs of delayed and cancelled complexes of flights. The model uses the scheduled and affected service time of particular complexes to determine their delays caused by disruption. During the last decade, a considerable attention has been given to proactive schedule recovery models as a possible approach to limit flight delays associated with Ground Delay Programs (GDP) (Abdelghany et al., 2004; Clarke, 1997). In these models, the impact of any reported flight delays, due to GDP or any other reason, is propagated in the network to determine any possible down-line disruptions (Monroe and Chu, 1995).

Wu (2005) explores the inherent delays of airline schedules resulting from limited buffer times and stochastic disruptions in airline operations. It is found that significant gaps exist between the real operating delays, the inherent delays (from simulation) and the zero-delay scenario. Results show that airline schedules must consider the stochasticity in daily operations. Schedules may become robust and reliable, only if buffer times are embedded and designed properly in airline schedules.

III. PROPOSED SYSTEM



Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results. If you know the decision tree algorithm. You might be thinking are we creating more number of decision trees and how can we create more number of decision trees. As all the calculation of nodes selection will be same for the same dataset. Yes. You are true. To model more number of decision trees to create the forest you are not going to use the same apache of

constructing the decision with information gain or gini index approach.

IV. CONCLUSION

From the analysis done on Flight data, The delay distribution were centered around zero. Some flights were delayed by more then 2 hours. December has largest amount of delays mainly due to snowstorms also June and July have delays due to summer vaccations. October and november are the months with least amount of delays. we see a marked "V" shaped decline in delay with the lowest delays in early morning hours. Both departure and arrival delays © 2017, IERJ All Rights Reserved Page 3 accumulate from the earlier morning hours reaching their peaks in the evening hours indicating Flight Delay Propagation.

REFERENCES

1 Abdelghany, K. F., Abdelghany, A. F., and Raina S., (2004) A model for projecting flight delays during irregular operation conditions, Journal of Air Transport Management, Volume 10, Issue 6, Pages 385-394

2 Aisling, R., and J.B. Kenneth, (1999) An assessment of the capacity and congestion levels at European airports, ERSA conference papers ersa 99, pages 241, European Regional Science Association.

3 Allan, S.S., S.G. Gaddy, and J.E. Evans, (2001) Delay Causality and Reduction at the New York City Airports Using Terminal Weather Information, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, Lexington, Massachusetts

4 Allison, P. D., (1999) Logistic Regression using the SAS System, John Wiley & Sons, Inc,

5 Ashford and Wright, (1992) Airport Engineering, John Wiley & Sons, Inc,

6 Bureau of Transportation Statistics, Airline On-Time Statistic. U.S. Department of Transportation. Washington, D.C. http://www.bts.gov/programs/airline_information.

7 Bracciali, C. F., X. Li, and A. S. Ranga, (2005) Real orthogonal polynomials in frequency analysis, Math. Comp. Volume 74, pages 341-362.

8 Christodoulou, C., and Georgiopoulos, M., Applications of Neural Networks in Electromagnetics, Artech House, Boston, 2001.

9 Daruis, L., O. Njåstad and W. Van Assche, Para-orthogonal polynomials in frequency 106 analysis, Rocky Mountain J. Math., volume 33, pages 629-645.

10 Girden, E. R., ANOVA: repeated measures, Newbury Park, Calif., Sage Publications, 1992.

11 Hagan, M., T., and Menhaj, M., (1994) Training feedforward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks, Volume 5, No. 6, pages 989-993.

12 Hansen, M., and C. Y. Hsiao (2005), Going South An Econometric Analysis of US Airline Flight Delays from 2000 to 2004, Presented at the 84rd Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2005.

13 Hansen, M., and D. Peterman, (2004) Throughput Impacts of Time-based Metering at Los Angeles International Airport, Presented at the 83rd Annual Meeting of the Transportation Research Board (TRB), Washington D.C., 2004.

14 Hansen, M., S. J. Tsao, A. Huang, and W. Wei, Empirical Analysis of Airport Capacity Enhancement Impacts: A Case Study of DFW Airport, presented at the 1999 Transportation Research Board Annual Meeting, Washington, D.C., 1999.

15 Hansen, M., (2002) Micro-level analysis of airport delay externalities using deterministic queuing models: a case study, Journal of Air Transport Management Volume 8, Issue 2 , Pages 73-87.

