International Journal of Emerging Technology and Innovative Engineering
Volume 4, Issue 5, August 2018 (ISSN: 2394 – 6598)

PREDICTION OF AIRLINE DELAYS USING K NEAREST NEIGHBOR ALGORITHM

G. Priyanka

Assistant Professor, Sri Krishna College of Technology

Coimbatore, India

ABSTRACT

According to the Bureau of Transportation Statistics, approximately twenty percent of the entire scheduled commercial flights are delayed. Airlines cause multi-billion dollars per year and cause great inconvenience to passengers. The primary goal of the model is to predict airline delays caused by inclement weather conditions using supervised machine learning algorithms. US domestic flight data and the weather data from 2013 to 2015 were extracted and used to train the model. K-Nearest Neighbors is implemented to build model which can predict delays of individual flights. In the prediction step flight schedule and weather data for the year 2016 were gathered and fed into the model. Using those data, the trained model performed a binary classification to predict whether a scheduled flight will be delayed or on-time.

Keywords—Airline delays, K Nearest Neighbor

I. INTRODUCTION

'Bureau of Transportation Statistics (BTS)' has provided information that approximately twenty percent of the entire scheduled commercial flights are delayed. BTS has categorized airline delays into five main causes, which are air carrier, extreme weather, National Aviation System, late-arriving aircraft and security. Weather is not only one of the main reasons of delays but it is also closely related to other categories, indeed. For example, National Aviation System category can include delays due to the re-routing of flights by inclement weather. Besides, weather is also a factor affecting late-arriving aircraft although airlines don't report the causes as weather. By considering those facts, weather's percentage share accounts for about 40% of total delay minutes. Thus, study on the influence of inclement weather on airline delays is essential for efficient flight operations. Furthermore, a decision support tool built on the study can inform passengers and airlines about weather induced delays in advance and help them reduce possible monetary losses. For this purpose, the classification model to predict weather-induced delays of individual flights is proposed in this work.

On-time performance of flights has been an important research subject as demands for air travel increase. Thus, several attempts were there to discover patterns in air traffic. In this project, arrival delays of individual flights using supervised machine learning algorithm is mainly focused. There are several reasons to explain why machine learning was tried. First of all, the volume of historical flight and weather data are too large to analyze analytically. Moreover, relationships between causal factors and delay or even correlations among factors are extremely complicated and highly nonlinear to test all hypotheses. Machine learning is able to develop models vigorously with huge amount of dataset and it has the ability to discover and display the hidden patterns in the data. To predict delays of individual flights, supervised machine learning algorithm is implemented with features including flight schedules and weather conditions at the origin and the destination.

II. RELATED WORK

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. The Federal Aviation Administration (FAA) describes the increase in delays and cancellations from 1995 through 1999.

Schaefer and Miller [14] 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. [15] 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). Hansen [16] analyzes runway delay externalities at Los Angeles International Airport (LAX) using a deterministic queuing model. The model allows estimating the delay impact of each specific arriving flight on each other specific arriving flight. The primary goal of the machine learning model is to predict airline delays. A flight is categorized as on-time if the departure delay is less than 15-minutes or delayed if the departure delay is more than 15-minutes.

III. PROPOSED SYSTEM

Airline delays cost airlines multi-billion dollars per year and cause a great inconvenience to passengers. Passengers experience increase in the time required for travel, experience inconvenience and stress, and may face additional expenses for food and lodging. Airline delays are caused by various delay factors such as air carrier, extreme weather, National Airspace System, late-arriving aircraft and security. When air transportation delays are eliminated or reduced it would make air travel more attractive and the demand for it would increase. The proposed system considers weather delay for predicting the flight delays. Weather delays are caused by various meteorological conditions such as temperature, wind, snowfall, precipitation at the airport. The proposed system uses K Nearest Neighbor technique to build the model for flight delay prediction.

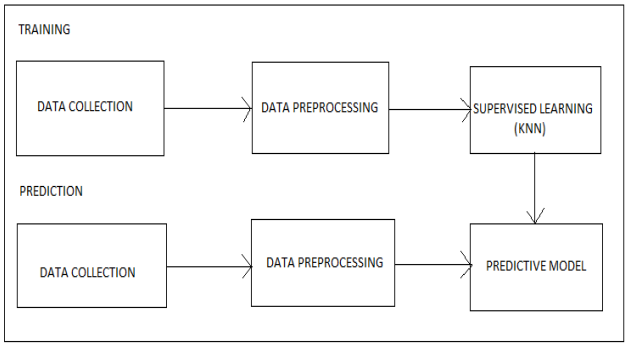

Fig 1: Architecture of proposed model

The above figure shows how the proposed machine learning model is built. Deploying a machine learning model includes the following steps namely data collection, data preprocessing, dividing data into training and testing sets, building a model on training data, evaluating the model on the test data. Finally if the performance is satisfying, deploy to the real system.

First step is data collection. At this stage, the relevant data has to be collected. Next step is data preprocessing. Data preprocessing involves data cleaning, data integration, data reduction and data transformation. Data cleaning is the process of detecting and correcting inaccurate data from a dataset. It also refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the coarse data. Data reduction is the transformation of numerical or alphabetical digital information derived empirically into a corrected form. The process involves the reduction of multitudinous amounts of data down to the meaningful parts. After the data has been cleaned and transformed it needs to be split into a training set and test set. This training data set is used to create the model which is used to predict the answers for new cases in which the answer or target is unknown. The model is trained with K Nearest Neighbor classifier. Once the model is built with the training data, it can be used to predict the targets for the test data. First the target values are removed from the test data set. The model is finally applied to the test dataset in order to predict the target values for the test data. The predicted value is then compared with the actual value. The accuracy of the model is defined as the percentage of correct predictions made. These accuracies can be used to compare the different models.

IV. IMPLEMENTATION DETAILS

1) Data collection

US domestic airline traffic data from 2013 to 2016 are obtained from the Bureau of Transportation Statistics (BTS)' Airline On-time Performance dataset. Weather data from 2013 to 2016 are obtained from the National Oceanic and Atmospheric Administration (NOAA)'s Integrated Surface Database. The BTS dataset contains on-time arrival performance data for non-stop domestic flights served by major air carriers. It also provides additional information such as origin/destination airports, flight numbers, flight schedules and delay times. NOAA's database contains weather information including wind, cloud height, visibility, temperature, pressure, precipitation, etc. reported approximately hourly at worldwide stations. The data is collected for a specific airline called jetblue airways with its origin at Boston International Airport (BOS) and destination at Los Angeles International Airport (LAX).

2) Data preprocessing

Data preprocessing involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing resolves such issues by preparing raw data for further processing.

Air traffic data for major airports and corresponding weather data are extracted. Following the rules of BTS, flights that arrive at the gate within 15 minutes of the scheduled time are considered as on-time and those that arrive at the gate exceeding 15 minutes of the scheduled time are considered as delayed. The following data fields were extracted from BTS dataset for every scheduled flight because those are factors having impacts on flight delays.

• Year
• Month
• Day of Month
• Departure and Arrival Schedule in Local Time
• Arrival Delay Indicator: 0 if actual arrival time minus scheduled arrival time is less than 15 minutes, 1 if actual arrival time minus scheduled arrival time is greater than or equal to 15 minutes.
• Departure Delay Indicator: 0 if actual departure time minus scheduled departure time is less than 15 minutes, 1 if actual departure time minus scheduled departure time is greater than or equal to 15 minutes.

It is known that delays are occurred in association with convective weather at the terminal area. Also, temperature conditions, high surface winds and precipitations make an aircraft landing difficult. Extracted weather fields reflecting these facts are as follows.

• Wind Speed Rate [*m/s*]
• Precipitation [*mm*]
• Snow Depth [*cm*]
• Temperature [F]

By preprocessing, all of categorical variables are converted to numerical variables since machine learning algorithms exhibit better performance with numerical variables.

3) Training and testing

A training set can be defined as a set of data which can be used to discover potentially predictive relationships. A test set is defined a set of data which can be used to assess the strength and utility of a predictive relationship. The US domestic flight data and the weather data from 2013 to 2015 are used to train the model. For testing the model flight data and weather data for 2016 are gathered and fed into the model.

4) Classification

Classification model was trained with flight data and weather data to predict arrival delays of individual scheduled flights. The classification model is typically trained on the dataset that contains details of the flights departing from Boston International Airport (BOS) and arriving at Los Angeles International Airport (LAX).

K Nearest Neighbor algorithm is used for implementing the model. The model is applied to the test data set to predict the expected values for the test data. The predicted value is then compared with the actual value.

K-Nearest-Neighbor Classifier (KNN)

The KNN algorithm is known as a non-parametric method which is mainly used for classification. The input to the KNN algorithm comprises of the k closest training examples in the feature space. The output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbours. K is a positive integer, typically small. For k = 1, the object is simply assigned to the class of that single nearest neighbor.

A step by step procedure on how to compute K-NN algorithm is given below.

1) Determine parameter k=number of nearest neighbours.

2) The distance between the query-instance and all the training samples must be calculated.

3) Sort the distance and determine nearest neighbours based on the K-th minimum distance.

4) Gather the category Y of the nearest neighbors.

5) Simple majority of the category of nearest neighbors is used as the prediction value of the query instance.

This algorithm predicts whether a flight will be on-time or delayed.

V. PERFORMANCE ANALYSIS

The performance metrics such as precision, recall and accuracy are used for predicting correctness of the results. Precision can be defined as the fraction of retrieved instances that are relevant. Recall can be defined as the fraction of relevant instances that are retrieved. Precision and recall can be calculated by the following formula

Precision = TP / (TP+FP)
Recall = TP / (TP + FN)

Total classes that are exactly correct are said as True Positive (TP). Classes that are predicted to be negative are called as False Negative (FN). Classes that are predicted to be positive are called as False Positive (FP).

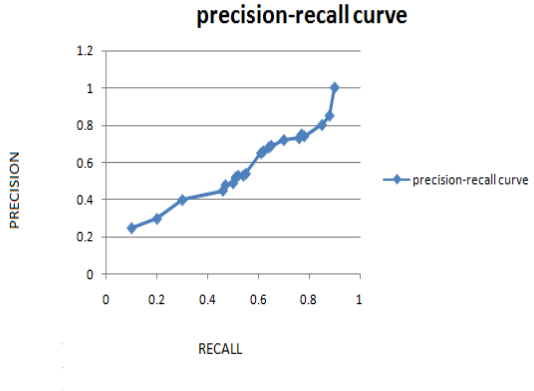

Fig 2: Precision-Recall curve

The above figure shows that the proposed work has results with 95% recall, 90% precision and 90% accuracy.

VI. CONCLUSION AND FUTURE WORK

This study proposed a prediction model to classify airline delays caused by inclement weather condition. In particular the model is built on historical weather and flight data by utilizing machine learning algorithms. K Nearest Neighbor which is a supervised machine learning algorithm is implemented in this study. The model's prediction performance on the test set is analyzed.

There are still possible approaches that can improve the model in the future. The model can be implemented using some other classification algorithms. Weather forecast data can be used instead of historical weather data. Other factors for predicting airline delays can also be considered for building the model.

REFERENCES

[1] U. D. of Transportation, "February 2016 on-time performance up from previous year," jan 2016.

[2] H. B. Juan Jose Rebollo, "Characterization and prediction of air traffic delays," Transportation Research Part C: Emerging Technologies, vol. 44, pp. 231–241, 2014.

[3] R. S. L. Alexander Klein, Chad Craun, "Airport delay prediction using weather-impacted traffic index (witi) model," in Proceedings of the Digital Avionics Systems Conference (DASC) 29th, 2010.

[4] Rong Yao, Wang Jiandong, Xu Tao, "Prediction Model and Algorithm of Delay Propagation Based on Integrated Consideration of Critical Flight Resources" in International Colloquium on Computing, Communication, Control and Management (ISECS), 2009.

[5] Sruti Oza, Somya Sharma, Hetal Sangoi, Rutuja Raut, V.C.Kotak, "Flight Delay Prediction System Using Weighted Multiple Regression" in International Journal of Engineering and Computer Science, vol 4,pp. 11765-11773, May 2015.

[6] Yanxiang Zhu, Nilesh Padwal, Mingxuan Li, "Data Analysis of U.S. Airlines Ontime Performance", vol. 5, pp. 55-60.

[7] Juan Jose Rebollo, Hamsa Balakrishnan, "A Network-Based Model for Predicting Air Traffic Delays" in 5th International Conference on Research in Air Transportation (ICRAT), vol. 20,no. 8, pp. 1034–1038.

[8] Yuqiong Bai, "Analysis of Aircraft Arrival Delay and Airport On-Time" Performance", M.S. Tongji University, 2006.

[9] D. A. Smith, "Decision Support Tool for predicting Aircraft arrival Rates from Weather forecasts," George Mason University, 2008.

[10] Shervin AhmadBeygi, Amy Cohn, Yihan Guan, "Analysis of the Potential for Delay Propagation in Passenger Aviation Flight Networks", University of Michigan, 2007.

[11] Klein, A., Kavoussi, S., Hickman, D., Simenauer, D., Phaneuf, M., and MacPhail, T., "Predicting Weather Impact on Air Traffic," *ICNS Conference*, Herndon, VA, May 2007.

[12] B. Sridhar, N. Chen, "Short term national airspace system delay prediction", Journal of Guidance, Control, and Dynamics, Vol. 32 No. 2, 2009.

[13] M. Jetzki, "The propagation of air transport delays in Europe", Thesis, Department of Airport and Air Transportation Research, Aachen University, 2009.

[14] Schaefer, L., and D. Miller, "Flight Delay Propagation Analysis with the Detailed Policy Assessment Tool", Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference, 2001.

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

[16] Hansen, M., "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, 2002.