Flight Delay Prediction using Machine Learning

Chetan Khobragade[1] Shashant Pandit[2] Devyani Shrikundwar[3] Abhishek Dahikar[4] Prof. Manisha Pise[5]

[1,2,3,4,5]Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, India

Abstract— Flight delays are quite frequent (19% of the INDIA domestic flights arrive more than 30 minutes late), and are a major source of frustration and cost for the passengers. As we will see, some flights are more frequently delayed than others, and there is an interest in providing the information to travelers. As delays are a stochastic phenomenon, it is interesting to study their entire probability distribution, instead of looking for an average value. This master's thesis proposes models to estimate delay probability distribution, based on a method called kernel density estimation and its extensions. These are data-driven methods, meaning that it does not try to model the underlying processes, but only consider past observations. Our models of increasing complexity have been implemented, optimized and evaluated on a large scale, using several years of records of INDIA domestic flights delays. During evaluation, we will measure the good performance of some of the models to predict delay distributions, in of the intrinsic difficulty of measuring the goodness of fit between a probability distribution and the corresponding random experiment.

Keywords: Flight Delay Prediction, Machine Learning

I. INTRODUCTION

The goal of this Machine Learning model is to predict whether the flight you are booking is likely to arrive on time or not. If you recognize that the flight is likely to be late, you would possibly prefer to book another flight according the needs. Moreover, the goal of this project is to estimate the probability of any flights to be more than x minutes late, for any x being the difference between the total connection time and the time to go to the departure gate. We would like to give this information to the customer during the search and reservation process, the model will have to give long-term predictions, up to several months forward and will not take into account short-term effects, like current weather or traffic situation. This model will be based on the unique public large dataset of flight delays. This dataset is only composed of domestic flights, with data from all major airlines.

Historically in charge of transaction processing for the travel and tourism industry, is developing new products to enhance the customer experience during the process of searching for a trip. In addition to the list of possible flight connections for a journey, a piece of information that can be provided is the risk of missing a connection. Knowing this probability can help the traveler to choose the best route, and the travel agent to adapt its suggestions or even prices.

In order to evaluate the risk of missing a connection, we need like to know the probability of the incoming flight being too late to be able to catch the second flight, taking into account the incompressible time necessary to go from the arrival gate to the departure gate of the second flight (possibly including immigration control). Models already exist to estimate the gate-to-gate transfer time. The goal of this master thesis is to build a model for the prediction of flight arrival delays. The prediction of short-term delays (for the next hours or so) is already a largely explored field. Indeed, using information about weather conditions, airports congestion and current flight delays allows quite accurate prediction of future delays, as some parameters influencing them are known, even if they still have component. For example, the website Flight-Caster exploit several sources of information (airports, airlines, weather and possibly historical data) to provide probabilities of being on-time, less than site is using the same estimations for all the flights when no short-term information is available.

We will first introduce the concepts and methods used during my thesis, covering both the model constructions and their evaluation. Then we will present the dataset used for this project, and we will try to identify the main factors influencing the delays.

Finally, we will present different methods to measure the performances of the models. These methods will be first used to optimize the models, in order to get the best possible predictions, and then will evaluate them in different use cases.

II. LITERATURE REVIEW

Flight delays hurt airlines, airports, and passengers. Their predictions is critical during the decision-making process for all the players of commercial aviation. Moreover, the development of accurate prediction models for flight delays became clumsy due to the complex system of air transport system, the different number of methods for prediction, and the deluge of flight data. By this context, this paper presents a thorough literature review of approaches used to build flight delay prediction models using the Data Science perspective. Also, we propose a taxonomy and summarize the initiatives used to address the flight delay prediction problem, according to scope, data, and computational methods, giving particular attention to an increased usage of machine learning methods. Despite of that, we also present a timeline of significant works that depicts relationships between flight delay prediction problems and research trends to address them.

In flight delay prediction models we can use various machine learning algorithm. Previously we can used the Support Vector Machine(SVM) this algorithm is that it has several key parameters that need to be set correctly to achieve the best classification results for any given problem. Parameters which will end in a superb classification accuracy for problem A, may result in a poor classification accuracy for problem B. The Random Forest (RF) algorithm is best for this model because it is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one among the foremost used algorithms, due to its simplicity and diversity (it is often used for both

classification and regression tasks). Also it is used for an ensemble classifier that consists of many decision trees, and outputs the mode of the classes output by individual trees. It combines the concept of bagging with the random selection of variables at each tree split. Benefits of Random Forests include the automatic generation of variable importance, their low sensitivity to outliers in the training data, and their good performance in cases where the number the variables is large compared to the number of samples. The RF approach has also been extended to regression method or technique.

III. DISCUSSION & CONCLUSION

Flight delays are the very crucial subject in the literature because of their economic and environmental impacts. They may increase the costs to customers and operational costs to the airlines. Apart from outcomes directly related to passengers or traveller, delay prediction is crucial during the decision-making process for every person in the air transportation system.

In this context, researchers created flight delay models for delay prediction over the last years, and this work contributes with an analysis of these models from a Data Science perspective. We developed a taxonomy scheme and classified models in respect of detailed components to clarify more.

Additionally, the taxonomy includes domain and Data Science branches. The former branch categorizes the problems (flight delay prediction) and the scopes. The last branch groups methods and data handling. It was seen that the flight delay prediction is classified into two main broad categories, such as delay propagation and root delay and cancellation. Besides, the scope determines one of the three specific extents: airline, airport, en-route airspace or an ensemble of them.

Apart from that, considering Data Science branch, we aimed at the datum, by categorizing data sources, dimensions that can be used in the models, and data management techniques to preprocess data and improve prediction models efficiency. We also studied and divided the main methods into three categories: Datasets, Data Cleaning, Probabilistic models, and machine learning. Those categories have been clustered as their use on specific forecast models for flight delays.

Among the taxonomic scheme, we also presented a timeline with all articles to spot trends and relationships involving the main elements in the taxonomy. In the light of the domain problem classification, this timeline showed a dominance of delay propagation and root delay over cancellation analysis. Researchers used to focus on statistical analysis and operational research approaches in the past. However, as the data volume grows, we noticed the use of machine learning and data management is increasing significantly.

REFERENCES

[1] https://en.wikipedia.org/wiki/Flightcancellation and delay
[2] https://www.kaggle.com/tylerx/flights-and-airports-data for Dataset.
[3] https://docs.microsoft.com/en-us/learn/modules/predict-flight-delays-with-python/
[4] M. Abdel-Aty, C. Lee, Y. Bai, X. Li, and M. Michalak. Detecting periodic patterns ofarrival delay. Journal of Air Transport Management, 13(6):355–361, Nov. 2007.
[5] K. F. Abdelghany, S. S. Shah, S. Raina, and A. F. Abdelghany. A model for projectingflight delays during irregular operation conditions. Journal of Air Transport Manage-ment, 10(6):385–394, Nov. 2004.
[6] S. AhmadBeygi, A. Cohn, Y. Guan, and P. Belobaba. Analysis of the potential fordelay propagation in passenger airline networks. Journal of Air Transport Management,14(5):221–236, Sept. 2008.
[7] S. Ahmadbeygi, A. Cohn, and M. Lapp. Decreasing airline delay propagation byre-allocating scheduled slack. IIE Transactions (Institute of Industrial Engineers),42(7):478–489, 2010.
[8] ANAC. Agˆencia Nacional de Aviac¸˜ao Civil. Technical report, http://www.anac.gov.br/,2017.
[9] https://en.wikipedia.org/wiki/Flightcancellation and delay
[10] C. Ariyawansa and A. Aponso. Review on state of art data mining and machine learn-ing techniques for intelligent Airport systems. In Proceedings of 2016 InternationalConference on Information Management, ICIM 2016, pages 134–138, 2016.
[11] F. Azadian, A. E. Murat, and R. B. Chinnam. Dynamic routing of time-sensitive air cargousing real-time information. Transportation Research Part E: Logistics and Transporta-tion Review, 48(1):355–372, Jan. 2012.
[12] Juan Jose Robollo and Hamsa Balakrishnan ”Characterization and Prediction of Air Traffic Delays”,
[13] Jianmo Ni, Xinyuan Wang, Ziliang Li ”Flight Delay Prediction using Temporal and Geographical Information”, https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a032.pdf
[14] Brett Naul ”Airline Departure Delay Prediction”, Ng, “Cs229: Machine learning lecture notes,” Standford University Lecture, 2011.

