Literature Review: Big Data Analytics for Predicting Airline Delays

Group Member: Hengchuan Zhang, Robin Sah, Li Guan

1. Introduction

Delays are common on all types of transport vehicles. However, airline delays differ from other delays due to differences in the geographical distance traveled. Furthermore, due to the limitation on the number of airports, many people need to take a second transport vehicle to arrive at their final destinations. If airline delays happen, the chain effect is that many people are unable to reach their destinations on the same day or even within days. The cost of time and money due to airline delays is many times that of other forms of transportation. Therefore, our team will be focusing on reducing the chances of people taking potentially delayed flights.

2. Challenges

One of the challenges we meet is the data collection. We need data to train our model to predict airline delays, which ideally be historical data that contains all the airline information in the past few years. Also, we need to know which information is needed before cleaning the data. Finding the right data set and cleaning it properly will be a big challenge to us. We also need real time data to examine our model. Those real time data need to be collected on our own. The tool we use to collect those real time data will become another challenge.

The next challenge is machine learning. None of us in our team has experience in machine learning. We need to do a lot of research on which models we should use and how to use them within a semester.

3. Approaches

The paper “Airline Flight Delay Prediction Using Machine Learning Models” provides us a solution on where and how to collect historical data sets. The author narrows down the scale to JFK airport and takes the historical data from 2018 - 2019 on kaggle. (Tang 2021) In addition to that, the author also points out which information is valuable for airline delay prediction, such as day of the week, departure time, arrival time, distance of the fight, wind direction, temperature, etc. (Tang 2021) By knowing that information, we can clean our data set and prepare to train our machine learning model. Tang is not the only person interested in predicting airline delay. In the paper “Flight Delay Prediction Based on Aviation Big Data and Machine Learning”, Gui and his team extend that information in more detail. Gui and his team add the destination airport weather information as another factor to consider because sometimes airplanes wait at the airport due to the destination weather condition. (Gui 2020) With the information from these two papers, our team can focus on collecting data from the time of day, travel distance, air route and weather conditions.

Choosing which machine learning algorithm to use is vital in our project. Fortunately, both Tang and Gui’s team tested different machine learning algorithms in their paper. And their results show that the decision tree model is the most accurate model compared to others. In Tang’s paper, the decision tree model’s accuracy rate is 97%. (Tang 2021) In Gui and his team’s paper, the decision tree model’s accuracy rate is 90.2%. (Gui 2020) Their discovery is valuable to us, which means we can focus on using the decision tree model for our project.

We further examined the paper *Machine Learning Approach for Flight Departure Delay Prediction and Analysis* by Ehsan Esmaeilzadeh et. al. This paper discussed the types of delay and associated possibilities. By collecting the data from EWR, JFK, and LGA in 2018, the paper collected a total of 33,548 flights including both domestic and international destinations. Taking advantage of the MATLAB, they implemented the SVM models and achieved 79.95% accuracy. Compared with the decision tree model provided by Tang's paper, the SVM model is less accurate but provides its ability to include a larger scale of factors. Given the fact that flight delays and reroutes could be affected by hundreds of factors, the SVM model shows its flexibility. This paper provides a comprehensive insightful data collection and data processing techniques.

The paper *Airline delay prediction by machine learning algorithms by Khaksar et. al.* discussed various machine learning methods as well. Compared to the Ehsan paper, they collected a larger dataset and expanded their research scope from the United States to Iran. A total of 2,825,647 US dataset was processed by the machine learning models. We learned that by taking advantage of a hybrid machine learning method, we could increase the prediction accuracy. In the paper, they combined the traditional decision tree method with the cluster classification method, which boosted the accuracy by more than 10% for the US network and 5% for Iran network. This paper shows the feasibility and broad impact of our research.

Similarly, the research article "*Flight Disruption Insights with Big Data Analytics*" delves into using big data analytics and machine learning to predict flight delays and understand their impact on various aspects of air travel [4]. The author utilized structured data related to flights, airports, airlines, and weather information, aiming to pinpoint key factors that could potentially disrupt flight schedules and further analyze the impacts of these disruptions, particularly concerning connecting flights and the cascading delays that might ensue. The article presented challenges such as ensuring data quality and managing the complexity of integrating various datasets (like airline, flight, airport, and weather datasets) to discern factors influencing flight cancellations. The methodology was partitioned into three core phases: Data Engineering, Exploratory Data Analysis, and Prediction of Connecting Flight Disruptions. Through the application of machine learning techniques, specifically utilizing a model constructed through Linear Regression and Logistics Regression, the research aimed to classify and predict airline delays instigated by severe weather conditions. Despite achieving a preliminary 30% of positive prediction and navigating through the potential risk of overfitting the data with K-fold cross-validation method, the model elucidated the imperative of Regression Analysis in Machine Learning, Big Data Analytics, and highlighted the relevance of Cross Validation technique and Regularization in ML for crafting effective models. Furthermore, the article acknowledged the significant subsequent impact on connecting flights upon a delay exceeding 45 minutes and suggested potential enhancements for the model in future endeavors, potentially incorporating financial metrics related to delay-induced losses.

In the research article about "*Flight delay prediction based on deep learning and Levenberg-Marquart algorithm*" researchers looked closely at using a deep learning model, helped by the Levenberg-Marquardt (LM) algorithm, to guess flight delays accurately [3]. They made a model using something called stacked denoising autoencoders and checked how well it worked, focusing on finding a balance between how long the computer takes and the number of denoising autoencoders to make sure the model was just right. They had a big challenge with their data because it was imbalanced, meaning there were a lot more examples of non-delayed flights than delayed ones. They tried two different methods to balance the data in order to correct this, but ultimately decided to utilize under-sampling because up-sampling resulted in overfitting and took too long.

We can take a few lessons from this article and apply them to projects like ours, which uses Big Data to predict flight delays. In order for our model to function properly, it is first important that our data be balanced. We must be careful when doing this to ensure that our model can generate correct predictions. Then, by slightly altering the model, particularly with the use of useful algorithms like the LM algorithm, we can improve the accuracy of our predictions. The researchers demonstrated that the LM algorithm and denoising autoencoders improved the accuracy of their model's predictions of flight delays and provided a solid solution to this challenging issue.

4. Project Proposal

Our team will be focusing on predicting airline delays with the use of data of historical airline information to train the machine learning model. Then, we will use the prepared model to predict airline delays based on the real time data and store those data into our data warehouse. The expected final project should be deployed on a website. When the user types the start, destination airport and time, our website will access the data warehouse showing which flights are more likely to delay than others. The goal of our project is to help users avoid flights that may be delayed so that users can arrive at their destination on time.

Citation Page

[1] Yuemin Tang. 2021. Airline Flight Delay Prediction Using Machine Learning Models. In *2021 5th International Conference on E-Business and Internet (ICEBI 2021), October 15-17, 2021, Singapore, Singapore*. ACM, New York, NY, USA, 7 Pages. https://doi.org/10.1145/3497701.3497725

[2] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou and D. Zhao, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning," in IEEE Transactions on Vehicular Technology, vol. 69, no. 1, pp. 140-150, Jan. 2020, doi: 10.1109/TVT.2019.2954094

[3] Yazdi, M. F., Kamel, S. R., Chabok, S. J. M., & Kheirabadi, M. (2020). Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *Journal of Big Data*, *7*, 1-28.

[4] "Flight Disruption Insights with Big Data Analytics.", Krishna Bathula, https://activityinsight.pace.edu/ctappert/intellcont/KrishnaBathula-1.pdf

[5] Esmaeilzadeh, Ehsan & Mokhtarimousavi, Sajad. (2020). Machine Learning Approach for Flight Departure Delay Prediction and Analysis. Transportation Research Record: Journal of the Transportation Research Board. 2674. 036119812093001. 10.1177/0361198120930014.

[6] H. Khaksar; A. Sheikholeslami. "Airline delay prediction by machine learning algorithms". *Scientia Iranica*, 26, 5, 2019, 2689-2702. doi: 10.24200/sci.2017.20020