Small Components' Analysis and Flight Delay Prediction

Team 22: Ben Romdhane Nourchene/ Ben Hassen Sami/ Kanoun Firas/ Fessi Ali

Flight Routes Dataset

1 Introduction

In a more and more connected world, flight routes gave us a great framework in order to detect countries/cities that are socially isolated from the rest of the world.We will use Open Flight database[7]which contains more than more than 67000 routes and includes data about airports and airlines. Since we have already explored much of the network related with this data set during the first assignments we decided to focus on the components that were not seen.[6].

In section 2 we will analyze the data visually from the airport locations to the connections between them and try to visualize the countries having the heighest number of airports. In section 3 we will take a deeper look at the smaller components of the graph since we only focused on the giant one during the assignments. After discussing the above, we will take a look at an interesting country that is known for being anti social : North Korea. Finally, we'll finish this project by implementing a useful model to predict flight delays using some machine learning techniques and the 2015 flight delays and cancellations data-set[2].

2 Data Analysis

In the previous assignments, we didn't get the chance to visualise the data since we only explored it as connected graphs. This project is a great opportunity to explore and analyze it using maps so that we can exactly see how everything is connected.

Please do note that since we are using Folium, HTML does not render inside the notebook on github. Please run the notebook to get all the visualisations.

2.1 Airport location

We start off by visualizing all the airports across the countries. In figure**??**, we can see that first world countries are the ones who have the biggest number of airports. While bigger countries like Russia and Algeria have a lower number per area due to their large area.

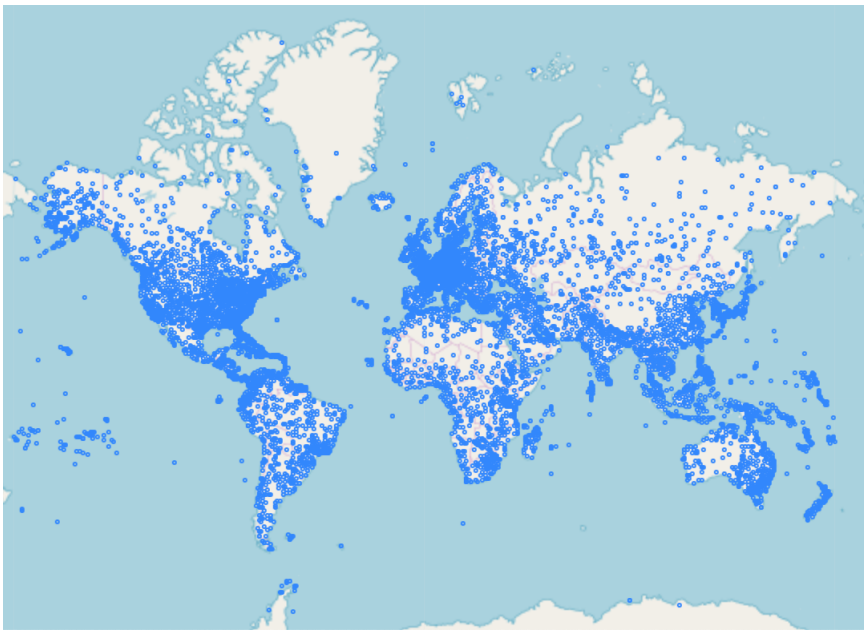

Figure 1: Airports locations around the world.

2.2 Airport Connections

Next, we take a look at how the airports are connected. Fig 2 displays all the edges of our graph on a map. We can notice that countries close to each other geographically have the most connections and that powerful countries are the most connected ones to the rest of the world. Indeed African and South American coutries are much less accessible than the European or the North American countries. This is due to the fact that these countries are more attractive for tourism as well as affairs, especially that they have a developed infrastructure to host a high volume of air traffic.

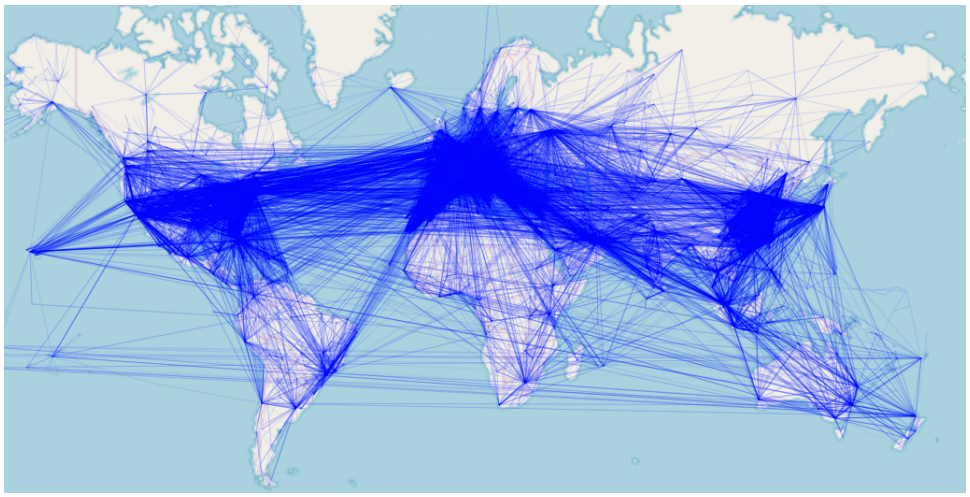

Figure 2: Connections between all the airports.

2.3 Countries with the most airports

After displaying each node and edge, a question one might ask is "Which are the countries that have the most airports?". In figure 3 we can see that the United States are leading the whole world in terms of number of airports.

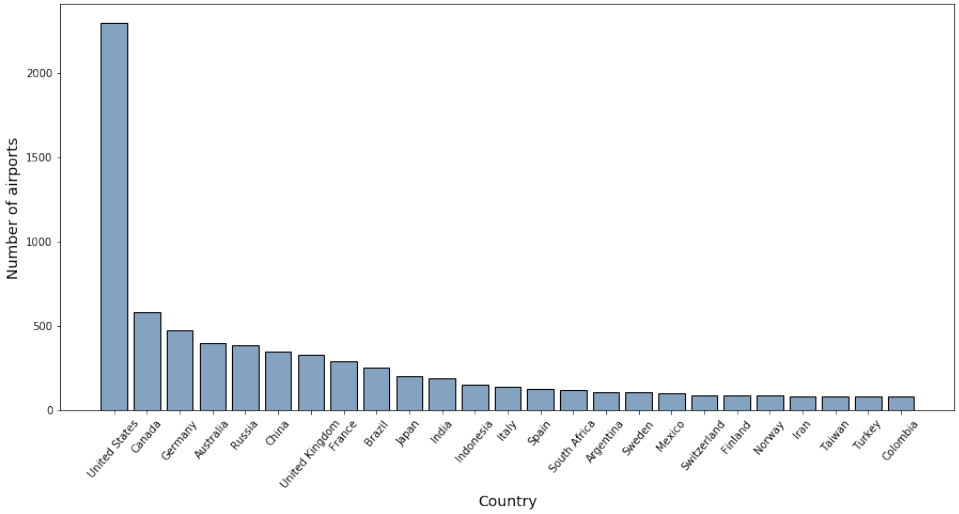

Figure 3: Countries with the most airports.

3 Smaller components analysis

Since all our work during the assignments was focused on the giant connected component, the question we kept asking ourselves was "How about the small components?" Therefore, we decided to locate these small components and try to understand the reasons behind their isolation from the rest of the world.Figure 4 is an example of the British Virgin Islands which is one of the 6 isolated component we found. While exploring these smaller components, we reached the conclusion that some countries might use local flights that are in

no way connected to other parts of the world due to many reasons, such as internal transports or the fact that some airports are intended for army such as in Namibia for example.

Figure 4: British Virgin Islands.

4 North Korea

As a result of its isolation, North Korea is sometimes known as the "hermit kingdom", a term that originally referred to the isolationism in the latter part of the Joseon Dynasty. Initially, North Korea had diplomatic ties with only other communist countries[5]. In this project we wanted to explore its connections with the outside countries.

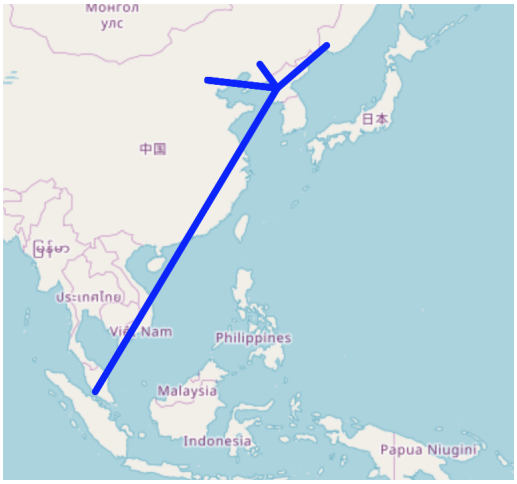

Figure 5: 1st Level North Korea Connections.

We can see from the figure **??** above that it is only connected to 3 countries :'Malaysia', 'China', 'Russia' through only one airport in each country except China where the number is 2. In the project's notebook we go even deeper and explore second level connections.

5 Predicting flight delays

In this section, we will explain the different steps we took in order to predict whether the delay corresponding to a flight departure time is negligible or not. We consider a delay negligible if the flight is only delayed by less than 30 minutes. Otherwise, we considered it as an important delay. After that, many models/classifiers were tried in order to come to the best results in terms of accuracy.

5.1 Data exploration

We have 7 features and 1 target in total :

- **AIRLINE**: A 2-Letter Airline Identifier.
- **ORIGIN_AIRPORT**: A 3-Letter abbreviation of the Starting Airport.
- **DESTINATION_AIRPORT**: A 3-Letter abbreviation of the Destination Airport.
- **SCHEDULED_DEPARTURE**: The Planned Departure Time.

- **DEPARTURE_TIME**: The exact time when the plane got its wheels off the track.
- **DEPARTURE_DELAY**: The Total Time Delay on Departure.
- **SCHEDULED_ARRIVAL**: The planned arrival time.
- **DELAY_LEVEL**: The target value we are trying to predict contains 1 if the delay is more than 30 minutes and 0 otherwise).

We then removed the columns that have information leakage (like departure time), and separated the target value.

5.2 ML Algorithms

We chose to look at 4 different classification models for our supervised learning task.

5.2.1 Gradient Boosted Trees

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.[1] This model was our first attempt at classifying the delay. We tuned the hyper-parameter n_estimators, which is the number of boosted trees we will use. We then built the optimal model using this parameter.

5.2.2 Random Forest Classifier

Another attempt was to use a Random Forest Classifier, an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees.Random decision forests correct for decision trees' habit of over-fitting to their training set[8]. We tuned the n_estimators parameter, which is the number of trees in the forest. Similarly, we used it to build the best possible model.

5.2.3 K Nearest Neighbors

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small)[3].

We needed to choose the number of neighbors to use for the knn algorithm. A tuning phase was used to find the best value and use it to build the optimal model.

5.2.4 Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.[4] We tuned the c parameter, which is the inverse of regularization strength. Smaller values of c specify stronger regularization. We then built our model based on this value.

5.3 Results

Algorithm	**Hyper-Parameter value**	**Accuracy**
Gradient Boosted Trees	17	87.021%
Random Forest Classifier	100	86.02%
K Nearest Neighbors	29	86.97%
Logistic Regression	0	87.02%

Table 1: Best Hyper-Parameters values and Model Accuracy

6 Conclusion

Through the course of this project, we explored the least connected components (Airports that are not connected to other parts of the world). We looked at a special case of a very well-known anti-social country "North Korea" and have seen hands on how disconnected it is from other parts of the world as it only allows a small number of flights from a few selected neighbouring countries (Malaysia, China and Russia) to come in. Finally, we looked for a more rich database with more information about flights such as the departure and arrival times to predict if a given flight would have a delay. In order to do this, we compared different methods of machine learning and tuned their hyper-parameters to get the best results we could possibly get with this data.

References

[1] Gradient boosting. https://en.wikipedia.org/wiki/Gradient_boosting.

[2] 2015 flight delays and cancellations. https://www.kaggle.com/usdot/flight-delays.

[3] k-nearest neighbors algorithm. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

[4] Logistic regression. https://www.statisticssolutions.com/what-is-logistic-regression/.

[5] North korea. https://en.wikipedia.org/wiki/North_Korea.

[6] Network tour of datascience github repository. https://github.com/mdeff/ntds_2018/tree/master/milestones.

[7] Airport, airline and route data. https://openflights.org/data.html.

[8] Random forest. https://en.wikipedia.org/wiki/Random_forest.