Flight Delay Prediction Using Graph Invariants

Thesis to be submitted in partial fulfillment of the requirements for the degree

of

Master of Technology

by

Mohammed Kashif M.
21MA60R39

Under the guidance of

Prof. Buddhananda Banerjee
and
Prof. Arindam Banerjee

DEPARTMENT OF MATHEMATICS

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Nov 2022

DECLARATION

I certify that

1. The work contained in this report has been done by me under the guidance of my supervisor.
2. The work has not been submitted to any other Institute for any degree or diploma.
3. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.
4. Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Place: Kharagpur
Date:

Department of Mathematics
Indian Institute of Technology,
Kharagpur
India - 721302

CERTIFICATE

This is to certify that we have examined the thesis entitled **Flight Delay Prediction Using Graph Invariants**, submitted by **Mohammed Kashif Mohammed Moizuddin**(Roll Number: *21MA60R39*) a postgraduate student of **Department of Mathematics** in partial fulfillment for the award of degree of Master of Technology in Computer Science and Data Processing. We hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance in partial fulfillment for the Post Graduate Degree for which it has been submitted. The thesis has fulfilled all the requirements as per the regulations of the Institute and has reached the standard needed for submission.

Supervisor

Department of Mathematics
Indian Institute of Technology, Kharagpur

Supervisor

Department of Mathematics
Indian Institute of Technology, Kharagpur

Place: Kharagpur
Date:

ACKNOWLEDGEMENTS

I take this opportunity to express my sincere thankfulness and deep regard to Prof. Buddhananda Banerjee and Prof. Arindam Banerjee, for the impeccable guidance, nurturing, and constant encouragement that they provided me during my post-graduate studies. Words seem insufficient to utter my gratitude to them for their supervision of my dissertation work. Working under them was an extremely knowledgeable experience for a student like me.

I also thank the department of Mathematics, IIT Kharagpur, and the Institute's Library for extending their support in many different ways in my urgent need.

Mohammed Kashif M.
IIT Kharagpur
Date:

ABSTRACT

Machine learning models with better statistical analysis provide confidence in our result. Here the real-world data of flight delay has been converted into simulated data to get our estimates. The feature engineering involved simulated data giving us the average arrival delay of flights for a particular day. The Regularity of a graph is one of the derived parameter as an input/independent variable. It is an algebraic term that encapsulates graph structural property to make our results more intuitive. Multiple linear regression model is fitted on this data. Statistical tests provide significance of this variable/parameters on a response variable (with confidence interval of 95% taken as usual). Further studies involve the use of different graph invariants for analysis to get better results.

These days deep learning models have been used to implement flight delay prediction to achieve better performance. But abstract information of what is learned by these model, make it highly noninterpretable. Fitted model with regression analysis gives us a highly interpretable model. Flight delay is inevitable and it plays an important role in both profits and losses of the airlines. An accurate estimation of flight delay is critical for airlines because the results can be applied to increase customer satisfaction and the incomes of airline agencies. A lot of work has been done on modeling and predicting flight delays, but most of them have tried to predict delays by extracting the main features and the most relevant features. However, most of the proposed methods are not accurate enough because of the massive volume of data, dependencies, and an extreme number of parameters.

Keywords: Regularity, Confidence interval, linear regression.

Contents

List of Figures

Chapter 1

Introduction

The Digital revolution just unlike any other revolution has changed everything around us. Nowadays, organizations of all types and across all industries utilize digital solutions to facilitate work operations, reduce costs, making error-free decisions. People worldwide are using technology to work, study, socialize, entertain, shop, or do online banking. On a common ground people making interactions with technology, generates a large amount of digital data, combined with the data produced by organizations worldwide creating a massive amount of digital data.

The availability of a large amount of data provides us with plenty of opportunities to utilize its insight for the sole purpose of marketing strategies, business development, and many more. The total data created, captured, copied, and consumed worldwide is forecasted to increase rapidly, reaching 59 zettabytes in 2020, According to Statista. Most data are created digitally and never find their way into papers.

Classification is a term that refers to predicting which given data points belong to which class. So, in this report, we are discussing linear regression, but later on, it will be implemented on different models. Based on performance, the final model, which is giving high performance, will be deployed. Machine learning models are very useful as many e-commerce companies, social media network platforms, websites, big tech giants, and many more organizations used to work on them to make decisions.

Chapter 2

What is Data?

Data is information about a person, facts, and statistics that may be qualitative or quantitative. This information of a person is useful or more specifically collectively for a large group of people from the dataset. Let the example of Employees in a company has their information related to their jobs such as employment Id, Age, Skills, joining date, and work profile. These data for faster processing are being converted into different file formats like CSV, Excel, Tabular, Text format, etc.

This data representation can be done in terms of any data structure, a pictorial representation that is in terms of graphs. Many machine learning algorithms employ various statistical, probabilistic, and optimization methods to learn from past known experiences. This learning finds useful patterns from large, unstructured, and complex datasets. These algorithms have a wide range of applications, in marketing just as customer purchase behavior detection, in the medical field for disease modeling, information theory for junk e-mail detection, and spam message detection. These applications are supervised learning variants rather than unsupervised learning. In the supervised learning method of Machine learning, decisions are made by the learning dataset. At the learning stage labeling of data is known and this information is used for making decisions on unlabeled examples. The following is the NOIR classification of data.

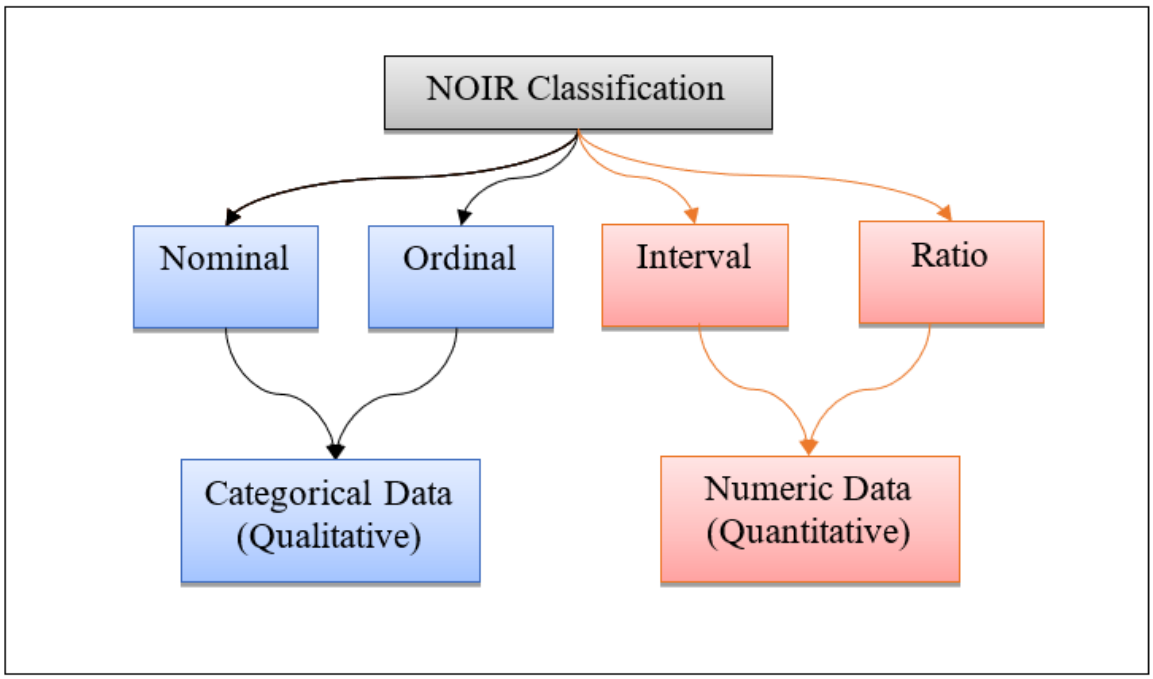

Figure 2.1: Data Classification

Chapter 3

Machine Learning

This is a topic under the recent application useful in many applications. As in recent years or decades, two major changes have boosted machine learning use in a practical scenario. These are,

(a) Advent of high-speed processors.

(b) Huge amount of data being generated

Machine learning is the study of computer algorithms in a way that it improves/teaches itself through experience and by the use of data. It has included some parts of computational statistics and is part of Artificial intelligence. The algorithms used in machine learning are applicable in various works such as email filtering, sentiment analysis, customer feedback support (Chat Box), speech and text recognition, and even in object detection techniques. In earlier days this application cannot fulfill the system requirement and was not feasible to develop such an algorithm to perform on computers. In recent processors, these tasks become easier, even though in recent days some machine learning algorithms are taking lots of time for their learning.

In nutshell, Machine learning more focuses on making a decision based on predictions made by computers. Not always machine learning statistics, but it also focuses on Data Mining. Data Mining is a field that has roots in exploratory data analysis. One part is machine learning i.e Deep learning which mimics

the biological brain. In the field of business analytics, machine learning is often referred to as predictive analysis

Machine learning uses programs in terms of algorithms that modify certain variables' present data structure depending on the data provided to it. Thereafter it performs certain useful tasks. For advanced tasks, it could be impossible to create an algorithm manually. From the practical point of view, it will no longer be effective for a person to develop such an algorithm.

Chapter 4

Linear regression

Linear regression is a machine learning algorithm based on supervised learning. It runs a regression task. Regression model targets a dependent variable based on independent variables. It is mainly used to find relationships between variables. Different regression models differ based on the type of relationship between dependent and independent variables under consideration and the number of independent variables used.

4.1 Simple Linear Regression:

Linear Regression performs the task of predicting the value of the dependent variable (y) based on a specified independent variable (x). Therefore, this regression technique finds a linear relationship between x (input) and y (output). hence the name linear regression.

$$y_i = \beta x_i + \alpha + \epsilon_i$$

4.1.1 Gradient Descent:

To update coefficient values to reduce the Cost function (minimizing RMSE value) and achieve the best fit line, the model uses Gradient Descent. The

best fit line for a given training dataset can be find using Gradient descent in a smaller number of iterations. The idea is to start with random b0 and b1 values and then iteratively update the values, reaching the minimum cost.

4.2 Multiple Linear Regression:

The following is a set of methods intended for regression in which the target value is expected to be a linear combination of the features. When there are multiple predictors, the equation of linear regression is simply extended to carry more variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \epsilon_i$$

Instead of a line, we now have a linear model, the relationship between each coefficient and its variable (feature) is linear.

Chapter 5

Experimental Details

5.1 Steps involved in the experiment:

(a) Simulate Data using Poisson distribution:
- Generate Poisson distribution with lambda = 750 (average number of lights per day)
- Perform an "iter" number of iterations, such that in each iteration, select the nth day, giving n[i] several flights.
- Randomly select n[i] a number of flights (data points) from the original data.

(b) Compute the following variables from the above simulated data:
- Average departure delay
- Average distance
- Maximum Indegree of graph
- Regularity of graph

(c) Fit the Multilinear regression model

(d) Analyse performance

Variables used for training are average departure delay, average distance, maximum Indegree of flight, and regularity of graph.

5.2 Variables under consideration:

(a) Average departure delay (avgDD):
Average departure delays of all flights scheduled for a given 24 hours

(b) Average distance (avgDIST):
Average arrival delays of all flights scheduled for a given 24 hours

(c) Maximum Indegree of flight (maxIndegree):
A maximum number of flight arrivals on that day at a particular airport.

(d) Regularity of graph (Regularity):
Regularity of graph is a graph theoretic invariant used to capture all geometric structural information. Macaulay2 is software used for regularity computation.

(e) Average arrival delay(response variable-avgAD):
Average arrival delays of all flights scheduled for a given 24 hours

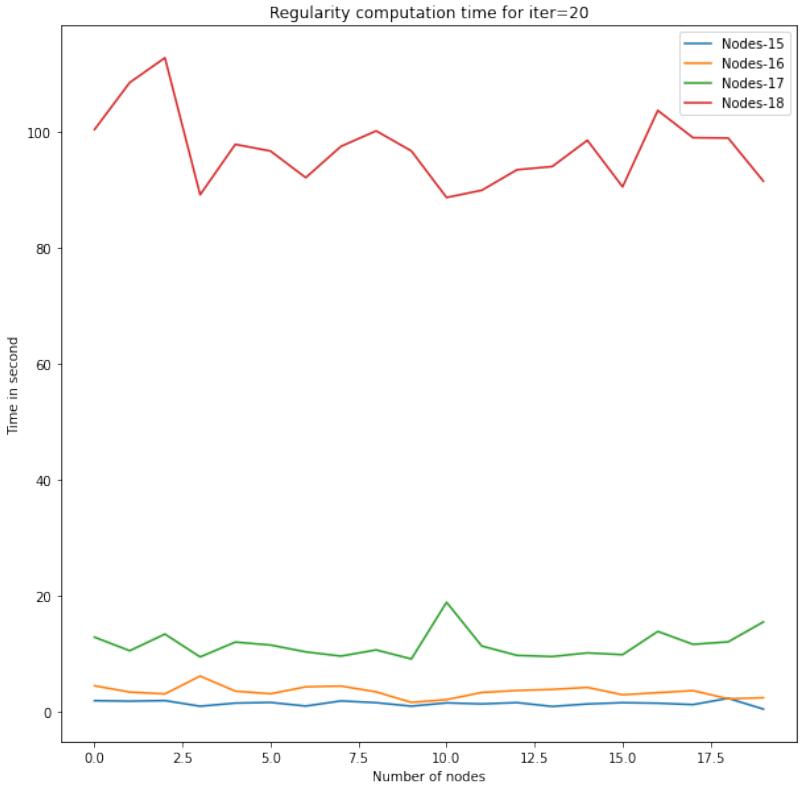

Figure 5.1: Computation time for Regularity

Regularity calculation is computationally very expensive as shown in Fig. 5.1. Computed for each day with graph having edge weight as the maximum number of flight arrival at that airports in a given day. As number of nodes increases computational time increases exponentially.

The Multiple linear regression model (with all variable as discuss in chapter 5) has residual term following the distribution as shown in Fig. 5.2. Thus, the assumption that require by hypothesis testing is followed.

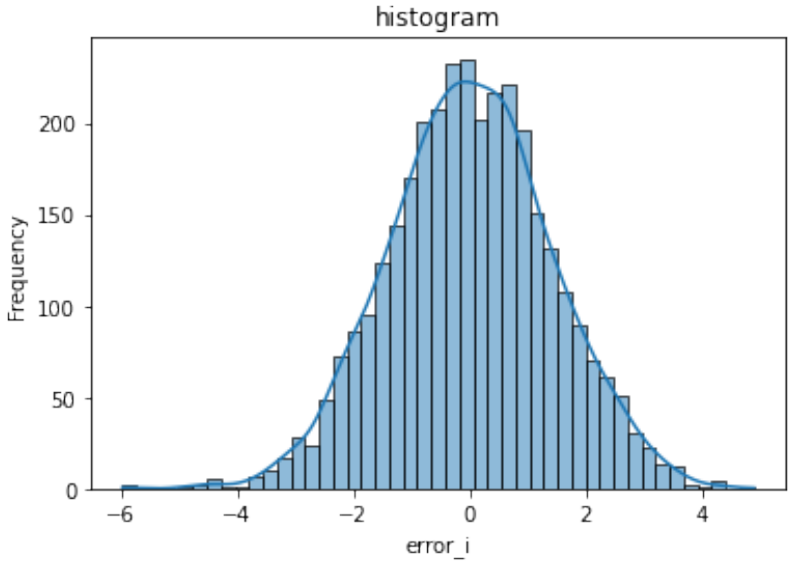

Figure 5.2: Normal distribution of residual term

5.3 Coefficient of Correlation and Partial correlation coefficient:

The R2 score was used earlier but it is coming out to be large. The additional used parameter does not lead to any significant improvement. Hence, this work is done by considering some other performance measures, like the coefficient of correlation and partial correlation coefficient. Finding relationships between variables in a dataset helps us to eliminate redundant variables for prediction, and how rich the dataset is explained by hypothesis testing.

One of the main issues rising concerns the dependency on different random variables. We can account for this issue using a correlation matrix, linear fit, and other different methods to find the relationship and use it in our model. But what if we have two variables that have a strong correlation between them but the truth is they are both tightly related to a third variable that increases the strength of this relationship?

This issue can undermine our assumptions and lead to incorrect results. One possible technique for explaining this problem is to use a measure called partial correlation. In contrast to the Pearson correlation, the partial correlation takes into account the presence and "control" over third variable.

Chapter 6

Observations

R2 score as performance measure:

avgDD	avgDIST	Regularity	MaxIndegree	R2-Score	p-value
1	1	0	0	0.9468	2.20e-16
1	1	1	0	0.9468	2.20e-16
1	1	0	1	0.9469	2.20e-16
1	1	1	1	0.947	2.20e-16

Table 6.1: R2 score including 'avgDD'

Here '1' means the input variable taken into consideration for R2 score computation and vice versa for '0'. While building our model one of our concerns is to use a minimum number of input variables, to reduce time complexity and make the model simple. We will try to find statistically insignificant input variables and avoid it from consideration. It has been observed that the R2 score is not changing considerably for different input variables.

	avgDD	avgDIST	maxIndegree	avgAD	Quantile_75	Quantile_90	Regularity
avgDD	1.000000	0.039157	-0.069628	0.945256	0.487892	0.688533	0.028210
avgDIST	0.039157	1.000000	-0.046245	0.005824	0.036027	0.042686	-0.132592
maxIndegree	-0.069628	-0.046245	1.000000	-0.071875	-0.070463	-0.082589	0.209821
avgAD	0.945256	0.005824	-0.071875	1.000000	0.603138	0.726938	0.042907
Quantile_75	0.487892	0.036027	-0.070463	0.603138	1.000000	0.563815	0.027227
Quantile_90	0.688533	0.042686	-0.082589	0.726938	0.563815	1.000000	0.026150
Regularity	0.028210	-0.132592	0.209821	0.042907	0.027227	0.026150	1.000000

Figure 6.1: Pairwise correlation coefficient (Pearson)

avgDIST	Regularity	MaxIndegree	R2-Score	p-value
1	0	0	0.9303	2.20e-16
1	1	0	0.9304	2.20e-16
1	0	1	0.9305	2.20e-16
1	1	1	0.9308	2.20e-16

Table 6.2: R2 score excluding 'avgDD'

As the above (Fig. 6.1) shows there is the highest correlation between the response variable 'avgAD' and input variable 'avgDD'. The above table shows the R2 score without using 'avgDD'. Still, the problem is not resolved. The other performance metric which is the partial correlation coefficient removes any dependencies between the input variable and shows the impact of a particular independent on the dependent variable. It has shown in the results section.

Chapter 7

Results

Variable	Partial Coeff. Corr.	p-value
avgDD	-0.330754115	9.46e-254
avgDIST	0.065209092	6.78e-11
Regularity	0.004533591	6.504109e-01
MaxIndegree	-0.025339279	1.129671e-02

Table 7.1: Partial corr. coeff. of predictor w.r.t response variable

$estimate

A matrix: 7 × 7 of type dbl

	avgDD	avgDIST	maxW	avgAD	Quantile_75	Quantile_90	Regularity
avgDD	1.000000000	0.11472670	-0.003835562	0.903611728	-0.330754115	0.079627711	-0.022070217
avgDIST	0.114726700	1.00000000	-0.014431402	-0.120993475	0.065209092	0.037178489	-0.123398353
maxW	-0.003835562	-0.01443140	1.000000000	-0.003197273	-0.025339279	-0.035500747	0.209911349
avgAD	0.903611728	-0.12099347	-0.003197273	1.000000000	0.436610182	0.178685208	0.032169834
Quantile_75	-0.330754115	0.06520909	-0.025339279	0.436610182	1.000000000	0.238958337	0.004533591
Quantile_90	0.079627711	0.03717849	-0.035500747	0.178685208	0.238958337	1.000000000	0.007745542
Regularity	-0.022070217	-0.12339835	0.209911349	0.032169834	0.004533591	0.007745542	1.000000000

Figure 7.1: Partial corr. coeff.

$p.value

A matrix: 7 × 7 of type dbl

	avgDD	avgDIST	maxW	avgAD	Quantile_75	Quantile_90	Regularity
avgDD	0.000000e+00	1.225195e-30	7.014130e-01	0.000000e+00	9.461657e-254	1.555649e-15	2.735162e-02
avgDIST	1.225195e-30	0.000000e+00	1.491123e-01	6.527958e-34	6.781624e-11	2.010133e-04	3.248292e-35
maxW	7.014130e-01	1.491123e-01	0.000000e+00	7.492653e-01	1.129671e-02	3.854331e-04	6.229807e-100
avgAD	0.000000e+00	6.527958e-34	7.492653e-01	0.000000e+00	0.000000e+00	1.719485e-72	1.297151e-03
Quantile_75	9.461657e-254	6.781624e-11	1.129671e-02	0.000000e+00	0.000000e+00	8.671462e-130	6.504109e-01
Quantile_90	1.555649e-15	2.010133e-04	3.854331e-04	1.719485e-72	8.671462e-130	0.000000e+00	4.387680e-01
Regularity	2.735162e-02	3.248292e-35	6.229807e-100	1.297151e-03	6.504109e-01	4.387680e-01	0.000000e+00

Figure 7.2: P-value for partial corr. coeff.

Chapter 8

Conclusion

(a) Average departure delay and average distance have a high correlation with average arrival delay.

(b) Regularity has the least correlation with average arrival delay. It is computationally expensive and has been tested to be statistically least significant.

(c) Max-Indegree is coming out to be a statistically significant input variable for this regression model.

Chapter 9

Future Work

(a) Experiments with the Regularity of the graph, with some improvement, would be employed for performance improvement.

(b) There is scope for further feature extraction from data with some simulation techniques, that will be studied and implemented in further work.

(c) To work on various methods of regression analysis to achieve better results.

(d) After regression analysis, the model will be built using different machine-learning techniques for further improvement.

Reference

1. https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018?select=2018.csv

2. Flight delay prediction based on deep learning and the Levenberg Marquart algorithm. (Maryam Farshchian Yazdi1 , Seyed Reza Kamel2* , Seyyed Javad Mahdavi Chabok2 and Maryam Kheirabadi)-Journal of Big Data.

3. Introduction to Linear Regression Analysis. (Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining)

4. Analyzing the Regularity of Complete k-Partite Graph using Super Strongly Perfect Graphs. (R. Mary Jeya Jothi, Ebin Ephrem Elavathingaln)- 2015 Online International Conference on Green Engineering and Technologies (IC-GET 2015)

5. Part (Semi Partial) and Partial Regression Coefficients Abdi, H. (2007). In N.J. Salkind (Ed.): Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage. pp. 736-740.