Regression Analysis using Machine Learning to Predict Flight Delays

Data Mining and Machine Learning PGDip in Data Analytics

Polina Prinii  
Postgraduate Diploma in Science in Data Analytics  
School of Computing, National College of Ireland, Mayor Street Lower, IFSC, Dublin1, D01Y300, Ireland Email: x21137757@student.ncirl.ie

*Abstract*—With the vast number of commercial flights available to the public both airports and airlines strive in the reduction of flight delays. Recent studies have examined multiple applications of various machine learning methods with the aim of predicting flight delays. Previous research has relied on conducting analysis based on a single route or airport conducted through complex methods such as Deep Neural Network. Using historical flight delay data recorded within the United States, the proposed research looks to identify if a relational strength is present between the y variable and the selected with the aim of determining with weather data is sufficient for the prediction of a delay. Contrary to the application of complex machine learning method, the proposed research successfully deemed that the application of simpler methods such as Linear Regression are just as effective in prediction. Overall regression methods applied in this research have demonstrated high accuracy scores indicating weather data is a strong indicator for delays within the commercial flights.

Keywords—machine learning, target, features, x, y, dependent variable, independent variable, regression.

# Introduction

Air travel has extensively grown since the flight of the first commercial flight from St. Petersburg, Florida to Tampa, Florida in 1914. Operating over 40 million flights a year pre the Covid-19 pandemic, commercial air travel has become a popular choice between the public. With the numbers of flights steadily increasing on a yearly basis many airlines as well as airports concern themselves with delays influencing their flights. Flight delays can cause adverse issues, from disgruntled customers to decrease in efficiency to an increase in capital costs. Airlines to optimise efficiency and reduce capital costs have turned to artificial intelligence, specifically machine learning analysis to predict future flight delays. The use of machine learning is not a novel idea, with multiple studies being completed on the use of machine learning algorithms to predict flight delays [1].

Since the introduction of meteorology weather forecasting has become a normal state of life. With weather forecasting readily available and reaching up to three months into the future, the proposed research looks to evaluate if weather factors contribute to a delay in a flight. Various weather factors such as extreme rainfall or winds, as well as well as concern of storms can be viewed as dangerous flying conditions thus ultimately delaying or even cancelling a commercial flight. For the proposed research, the study takes a simpler approach and evaluates metric such as precipitation, wind, and dry bulb temperature to determine if a prediction can be reached for future delays.

With the current battle against the Covid-19 pandemic the number of commercial flights reduced significantly since the confirmation of the first case back in January 2020. During the year of 2020 it was estimated that only 16.9 million flights operated within the year, reducing yearly flight by a half. With the virus running rampage a clear indicator emerged to the explanation of flight cancellations. However, in recent months the industry has seen a steady growth with over 25 million flights operated in 2022 [2].

As the commercial aviation industry steadily increases operations, concerns surrounding efficiency and capital costs resurface. It is to no surprise that the industry would turn to machine learning methods to predict flights delays. To successfully predict a delay the chosen data must be accurate and well trained to ensure performance is at its best. Taking into consideration the Covid-19 pandemic has been influencing the industry since the early days of 2020, it is believed that the most accurate data originates from 2019.

The proposed research looked to evaluate simpler methods of machine learning such as regression analysis. Regression analysis studies the relationship between two or more variables of interest. In the proposed research it was identified the target variable is the measure of delay. The targe variable or otherwise known as the y – dependent variable. The accuracy behind the chosen data is crucial as the proposed research sets out to evaluate weather predictors

The aim of the proposed research is the application of several machine learning regression algorithms to a chosen dataset which represents all recorded flight delays within the United States and the corresponding weather conditions for a given record. The proposed research aims to answer the following question:

*“Could the different weather variables be effective in predicting a delay or cancellation to a commercial flight using Machine Learning methods?”*

The relative success of each regression algorithm applied to the chosen dataset will be assessed and compared based on multiple indicators such as the R2 value. The relative success of each regression algorithm is dependent on feature selection which ultimately ensures that the best feature variables (x variables) are selected to weigh against the target variable (y variable).

The proposed research hopes to comprehensively evaluate the machine learning algorithms which have been selected and applied to the chosen dataset to determine if the selected feature variables in the form of weather statistics are strong predictors for the prediction of future delays in commercial flights. Based on the results under covered, the proposed research looks to draft recommendations for future work which may look to evaluate various other factors such as airport efficiency through the application of the evaluated machine learning algorithms.

The following proposed research is undertaken utilising the open-source software Python. All supporting materials of the proposed researched have been uploaded to Github for the publics reference [3].

# Related Work

The prediction of flight delays through machine learning application is not a novelty idea, many have researched and published literatures exploring various machine learning techniques. A favorite amongst the machine learning application is the analysis of data and prediction of delays based on deep learning methods, a subfield of machine learning. Specifically, many literatures cover the application and evaluation of Deep Neural Networks to predict future flight delays [4][5][6][7][8], however some have taken to comparing the more complex methods from Deep Neural Networks against robust but less complex methods such as regression.

The application of Deep Neural Network is a complex one. Neural networks are layers of nodes which were designed to mimics the human brain specifically the neurons which make up the brain. These nodes sit within layers and are connected to adjacent layers. The complexity rises as more layers are introduced with the learning becoming deeper. In comparison to simpler robust methods such as Linear Regression, the development of a machine learning model using Deep Neural Networks can be viewed as a complex task. A downside to the application of Deep Neural Networks can lay with the understanding of how the outputted result is arrived at due to the vast number of components involved.

Many literatures have analysed similar feature variables to predict the identified target variable. These studies have achieved high accuracy scores, averaging at around 95% when considering all cited literatures. The proposed research aims to achieve such results upon the application of both unsupervised and supervised learnings.

The proposed research looks to unsupervised learning to identify underlying patterns in unlabeled data. Within the scope of the proposed research, the application of K-means clustering is conducted to classify flight delays, this method was selected due to its simplicity, rapid operation, and consequent applicability to large datasets. From previous research K-means clustering is used in conjunction with Decision Tree analysis [1]. For the proposed research the K-means clustering is used to identify patters and or structures within the chosen data as the regression algorithm selected for the research may not be compatible with K-means clustering. Further research has demonstrated that the application of K-means clustering can assist in defining the spatial relationship when working with Deep Neural Network to predict airport delays based on spatiotemporal analysis [9]. The proposed research hopes to achieve a high accuracy score as the chosen dataset is being analysed.

However, the proposed research is prepared to evaluate poor results due to the concern that K-means clustering is a much powerful tool when used for classification purposes. It is important to note that the main aim of the proposed research is regression analysis with the goal of predicting future flight delays.

The use of supervised learnings methods is considered just as popular and powerful as unsupervised when looking to address a research question. As discussed, many when addressing flight delay predictions undertake a mix of deep learning methods and simple yet robust machine learning methods such as regression models. It has been discovered that based on the selection of the data, selection of the features and target variables as well as the training and testing of the built models that the simpler methods can outperform the much more complex ones [10]. With the application of Random Forest regression outperforming the application of Artificial Neural Network by a major 24%. The proposed research will look to evaluate the Random Forest regression method with the means of comparison between two other supervised learning methods.

Nonetheless, this is not to deem the application of simpler methods as outperforming that of the complex ones. Another study has discovered the application of Neural Network Classifiers contains a higher accuracy score of that of Decision Tree and Logistic Regression analysis [11].

In addition to Random Forest application, the proposed research looks to apply Linear Regression and K Nearest Neighbor Regression with the aim of benchmarking the three against one another. Thus far, it has been noticed that flight delay prediction is evaluated through the primary application of deep learning models which are mostly supported by simpler methods such as Linear Regression [12]. Considering this pattern, the proposed research hopes to steer away from this pattern and to produce an easily comprehensible study which compares three different methods.

Similar work to the proposed research has identified that application of the selected machine learning methods can yield high successful results, though the feature selection utilised in the said studies, the proposed study believes that similar results can be achieved with the choice of feature variables against the target variables. The use of Gradient Boosting Classifier has demonstrated that a high accuracy score of 85.73 can be achieved [13]. Furthermore, the use of Multiple Linear Regression, which is equivalent to Linear Regression however, accounting for multiple independent variables influencing the chosen dependent variable can deduce high R2 values of 84% [14] when analysis the relationship.

Lastly, the proposed research is confident high-performance scores can be achieved for the chosen dataset considering review of similar research where K-Nearest Neighbor regression algorithm have been applied to predict flight delays based on various variables. It was discovered that the K-Nearest Neighbor method is capable of scoring performance scores between 70% and 80% [15]. A positive note to the proposed research prior to the discussion of results. In conclusion the proposed research notes that the use of flight data from the United States is of a common practice, as the proposed research has noted that the majority number of literature reviews used for the discovery and support of the proposed research use flight delay data originating from the United States. With the Federal Aviation Administration operation more than 45,000 flights a day, equating to almost half of the global number of flights, coming in at roughly 16.5 million flights a year [15].

# Methodology

The proposed research has chosen to follow the Knowledge Discovery in Databases otherwise known as the KDD process as illustrated by Fig1.

Graphical user interface, text, application, email, website

Description automatically generated

1. Knowledge Discovery in Database Process, containing a total of 6 sequntial steps.

The proposed research justifies its choice for the KDD process due to its sophisticated data mining technique to identify and evaluate patterns from data. The KDD process is widely used for machine learning, database management and even artificial intelligence application. However, the proposed study notes the KDD process is not a fixed sequential process and various stages of the process can be re-visited which will allow for the maximum extraction of knowledge from the chosen dataset. The following actions were undertaken under the guidance of the KDD process:

## Data Selection:

.

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

Identify applicable funding agency here. If none, delete this text box.

* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

1. Khaksar, H., Sheikholeslami, A. (2019). 'Airline delay prediction by machine learning algorithms', Scientia Iranica, 26(5), pp. 2689-2702. doi: 10.24200/sci.2017.20020
2. Statista, “Number of flights performed by the global airline industry from 2004 to 2022”, Accessed on April 30th, [Online], Available at: <https://www.statista.com/statistics/564769/airline-industry-number-of-flights/>
3. Github, “Data-Mining-Machine-Learning-Final-Project”, [Online], Available at: <https://github.com/polinaprinii/Data-Mining-Machine-Learning-Final-Project>
4. Y. J. Kim, S. Choi, S. Briceno and D. Mavris, "A deep learning approach to flight delay prediction," 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), 2016, pp. 1-6, doi: 10.1109/DASC.2016.7778092.
5. Yazdi, M.F., Kamel, S.R., Chabok, S.J.M. *et al.* Flight delay prediction based on deep learning and Levenberg-Marquart algorithm. *J Big Data* **7,**106 (2020). <https://doi.org/10.1186/s40537-020-00380-z>
6. K. Cai, Y. Li, Y. -P. Fang and Y. Zhu, "A Deep Learning Approach for Flight Delay Prediction Through Time-Evolving Graphs," in IEEE Transactions on Intelligent Transportation Systems, doi: 10.1109/TITS.2021.3103502.
7. Y. Jiang, Y. Liu, D. Liu and H. Song, "Applying Machine Learning to Aviation Big Data for Flight Delay Prediction," 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 2020, pp. 665-672, doi: 10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00114.
8. 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.
9. H. Zhang, C. Song, H. Wang, C. Xu and J. Guo, "Airport Delay Prediction Based on Spatiotemporal Analysis and Bi-LSTM Sequence Learning," *2019 Chinese Automation Congress (CAC)*, 2019, pp. 5080-5085, doi: 10.1109/CAC48633.2019.8996754.
10. S. Kaewunruen, J. Sresakoolchai, and Y. Xiang, “Identification of Weather Influences on Flight Punctuality Using Machine Learning Approach,” *Climate*, vol. 9, no. 8, p. 127, Aug. 2021, doi: 10.3390/cli9080127.
11. Kuhn, Nathalie, and Navaneeth Jamadagni. "Application of machine learning algorithms to predict flight arrival delays." *CS229* (2017).
12. Bin Yu, Zhen Guo, Sobhan Asian, Huaizhu Wang, Gang Chen, Flight delay prediction for commercial air transport: A deep learning approach, Transportation Research Part E: Logistics and Transportation Review, Volume 125, 2019, Pages 203-221, ISSN 1366-5545, <https://doi.org/10.1016/j.tre.2019.03.013>.
13. N. Chakrabarty, "A Data Mining Approach to Flight Arrival Delay Prediction for American Airlines," 2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON), 2019, pp. 102-107, doi: 10.1109/IEMECONX.2019.8876970.
14. Kalliguddi, Anish & Leboulluec, Aera. (2017). Predictive Modeling of Aircraft Flight Delay. Universal Journal of Management. 5. 485-491. 10.13189/ujm.2017.051003.
15. Mehmet Cemal ATLIOĞLU, Mustafa BOLAT, Murat ŞAHİN, Volkan TUNALI, Deniz KILINÇ; (2020), Supervised Learning Approaches to Flight Delay Prediction. Sakarya  
    University Journal of Science, 24(6), 1223-1231, DOI: <https://doi.org/10.16984/saufenbilder.710107>
16. Federal Aviation Administration, ‘Air Traffic by the Numbers’, Accessed on April 30th, [Online], Available at: <https://www.faa.gov/air_traffic/by_the_numbers/>
17. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
18. K. Elissa, “Title of paper if known,” unpublished.
19. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
20. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
21. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.