National College of Ireland

Project Submission Sheet – 2022/2023

Student Name: Polina Prinii ……...............………………………………………………………………………

Student ID: x21137757……………………………………………………………………………………………

Programme: PGDDA…………………………………………………………............. **Year:** 2022…………..............

Module: Data Mining and Machine Learning 1……………………………………………..

Lecturer: Barry Haycock……………………………………………..…

Submission Due Date: 1st of May ………………………………………………………………..

Project Title: Regression Analysis using Machine Learning to Predict Flight Delays…………………...........................

Word Count: 4300……………………………………………………………………………………………………..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Polina Prinii
……

Date: 1st of May 2022
……

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

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.

I. 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 R^2 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].

II. 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 R^2 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 [16].

III. METHODOLOGY

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

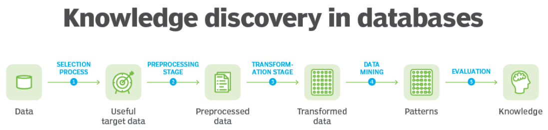

Fig. 1. Knowledge Discovery in Database Process, containing a total of 6 sequntial steps.

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:

A. *Data Selection:*

Model strength of a given machine learning algorithm heavily depends on data selection. Initially the proposed researched looked to work with three large and relatively similar datasets however, due to machine capabilities the proposed research had to significantly reduce its scope and ultimately one dataset was created for this research which met the requirements of the set-out work. The main was acquired from Kaggle [17] which describes all recorded flight delays and or cancellation for all 365 airports within the United States, alongside the weather statistics for each record.

To ensure the proposed research remained within the confines of the set-out scope of work, the proposed research undertook the following steps to draft the final dataset in preparation for pre-processing:

1. Due to the dataset from Kaggle being split into multiple .csv (Comma Separated Values) files, data was imported based on seasonality to ensure an accurate representation of weather indicators.

2. Random selection from the three imported datasets were made, with a 10,000 sample from each set.

3. The three randomly selected samples were merged into one final dataset.

4. Lastly, the final dataset was exported to a .csv file.

B. *Pre-Processing*

Pre-Processing is the activity of data cleansing and manipulation of the selected raw data. This step address issues such as null values and outliers. Following the selection of the final dataset, the dataset was found to have many null values as illustrated by Fig.2. Missing values primarily were presented within the '*actual_departure_dt'* and the *'actual_arrival_dt'* columns, reaching near to five hundred values per columns. These null values were address by setting a set date range from which the supporting code material populated the missing values for both columns.

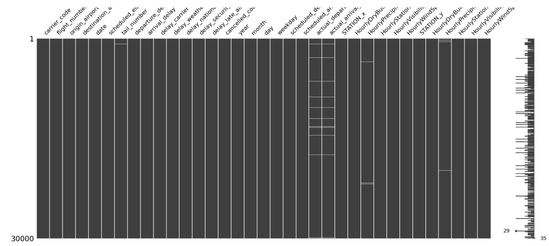

Fig. 2. Visual representation of all null/missing values within the dataset.

Additionally, all date type objects were converted to ordinal numerical values to allow for the selected regression algorithms to consider features representing date-time.

Lastly the proposed research made the conscious choice to not address any outliers. Due to the nature of the selected datasets the proposed research rationalised that the present features within the dataset may not specifically contain outliers, especially when referring to weather indicators as no one day is the same in terms of weather. Thus, the supporting code material plotted box plots to identify outliers however, no further steps were undertaken to address identified outliers.

All datasets prior to the application of the selected machine learning algorithm were split between train and test data following a 70/30 ratio. As the proposed research primarily worked with numeric value, all values were standardized prior to the application of the machine learning algorithm.

C. *Transformation:*

The transformation step ensures the data is prepared and developed for the data mining phase. Preparation may include feature selection and or dimensionality reduction.

The prosed research following the understanding that regression algorithms operate only on numeric value within the Sklearn package undertook feature selection and dimensionality reduction which was later utilised by the K-means clustering. Feature selection was determined using a correlation matrix, with any features being selected with a correlation above .25, Fig.3. illustrates the correlation matrix.

Fig. 3. Correlation Marix.

A total of thirteen features were selected from the overall thirty. The proposed research's feature selection is supported by the Information Gain which evaluates the gain of each variable in the context of the target variable.

Having completed feature selection, the proposed research performed Principal Component Analysis, a popular technique for dimensionality reduction in preparation for K-means clustering.

D. Data Mining:

In the Data Mining stage machine learning algorithms are applied, parameters altered, and results evaluated. The derived results will feed into the Evaluation stage of the KDD process. The following unsupervised and supervised learning methods has been applied to the chosen dataset.

a) Unsupervised:

K-means clustering, a popular clustering technique which can identify underlying patterns within the data. This technique was applied twice within the proposed study. Once prior to Principal Component Analysis and secondly after the selected features were dimensionally reduced to 2 features using Principal Component Analysis. Prior to the application of the K-means clustering method all data was standardized using StandardScaler() is not done so previously.

b) Supervised:

A total of four supervised learning methods were applied and evaluated against the chosen dataset.

The proposed research began with the Multiple Linear Regression method, as mentioned previously Multiple Linear Regression is identical to Linear Regression with the only difference being multiple independent variables are being proposed rather than a singular one. The Linear Regression model was trained and tested using the 70/30 split, with the target aka the y variable selected as '*departure_delay'* and a total of twelve features aka x variables.

Next the proposed research moved to K-Nearest Neighbor Regression application. Prior to running the proposed research determined the optimum k value, which represent the *neighborhood.* The KNN regression is a non-parametric method which in an intuitive manner approximates the association between independent and dependent variables by averaging the observations within the same *neighborhood.* Once the optimum k value was determined using the training set of the data following the 70/30 train/test split the remaining 30% of the model was evaluated.

In addition to Linear Regression and K-Nearest Neighbor regression the proposed research turned to Random Forest regression. Random Forest regression uses ensemble learning methods. Ensemble learning is a technique which combines predictions from multiple machine learning algorithms to build a more accurate prediction that a single model. As with the two previous regression methods the model is trained and tested based on the 70/30 ratio.

Lastly, the proposed research applied a Boosting method which is a type of ensemble learning. Boosting methods combine a set of weak learning into a strong learner to minimize training errors. Having trained and tested the chosen data based on the 70/30 ratio tow boosting methods were directly compared. These methods are Adaptive boosting otherwise known as AdaBoost and Gradient boost. Here the proposed research looked to simple benchmark the two against one another for future work.

E. Interpretation/Evaluation:

As the name suggests the final stage of the KDD process is interpretation and evaluation of the Data Mining steps. The following report being outlined for the proposed research will discuss in detail all result within the "Evaluation" section.

IV. Evaluation

The proposed research now moves to the evaluation of results derived from the Data Mining step.

It was found that through the application of K-means no distinct clusters could be found for the dataset having underwent the application of Principal Component Analysis, a form of dimensionality reduction. Prior to the application of clusters, the raw data was plotted, these plots can help pre-determine if distinct cluster exist prior to K-means application. Fig.4. illustrates the data prior to the implementation of K-means clusters.

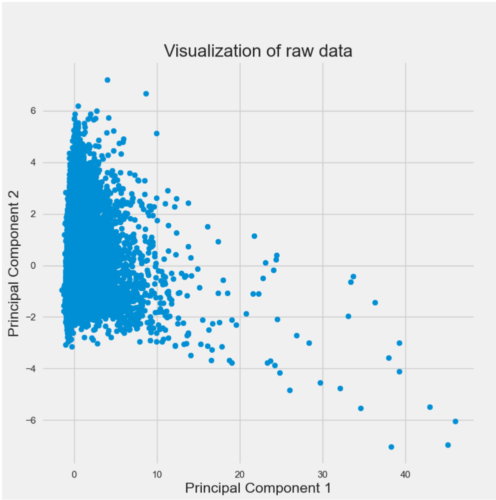

Fig. 4. Visualisation of Raw Data.

Fig.4. clearly demonstrates a lack of clear distinct clusters, a concern surrounding this arises as the application of supervised may yield poor results due to the extensive overlap as show above. However, considering the main application of the proposed research is regression analysis the concern of poor results is not alarming as K-means clustering is more effective for classification studies. Following the determination of the optimum k-value as illustrated by Fig.5.

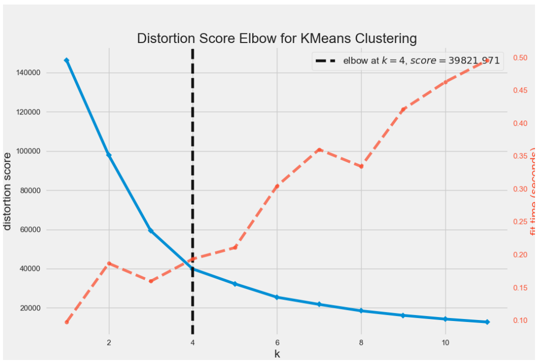

Fig. 5. Optimum K-Value.

the K-Means clustering method was successful in clustering the data as shown by Fig.6. however, the overlap persisted. This indicated that the selected features which have been dimensionally reduced following PCA application were not contributing factors within the Euclidean distance between the points.

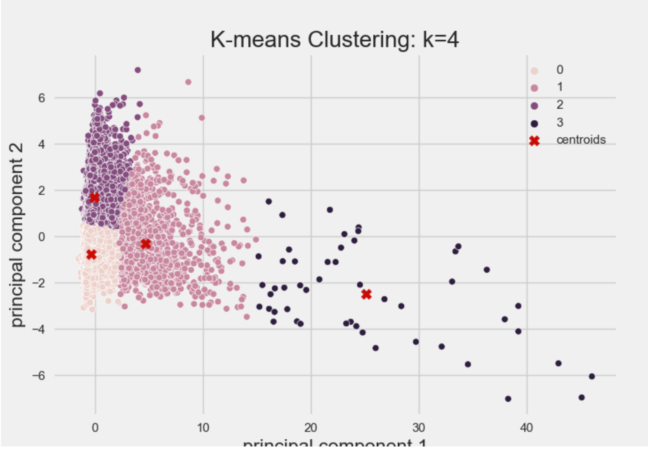

Fig. 6. K-means clustering of the PCA application data with the optimum k of 4 clusters.

The discovered result of K-means clustering was quite disappointing as the proposed research expected for the unsupervised learning method to cluster based on the origin airport. However, the proposed method served its purpose, and the proposed research took into consideration that the results of supervised learning methods may be impacted.

The selected supervised learning method are evaluated based on the following criteria:

1. R^2 value: The R^2 value is a statistical measure which represents the proportion of variance for a given dependent variable is explained by a single or multiple independent variables. The proposed research will look to regression models with the highest R^2 value.
2. Mean-Squared-Error (MSE): The MSE value is a measure of the amount of error within the model. MSE determines this error by assessing the average squared difference between the observed and predicted values. The proposed research will look to regression models with the lowest MSE value.
3. Root-Mean-Squared-Error (RMSE): The RMSE value is the standard deviation of residuals aka prediction errors. The RMSE measures how spread out these residuals are from one another. In other words, the RMSE tells how concentrated the data is around the regression line. The proposed research will look to regression models with the lowest RMSE value.
4. Mean-Absolute-Percentage-Error (MAPE): The MAPE value measures accuracy as a percentage. The proposed research will look to regression models with the lowest MAPE value.

As the overall aim of the proposed research is the application of simple yet robust machine learning algorithms to predict flight delays, the research deemed the above choice of evaluation methods as favorable due to the simplicity behind their understanding. In addition to the above evaluation methods, the proposed research turned to Ordinary Least Squares (OLS) regression, a technique which falls under supervised learning. This method operates by estimating the unknown parameters by creating a model which minimizes the sum of the squared errors between the observed data and the predicted, in this case between the train and test data.

The proposed research now moves to discussing the performance of each machine learning regression algorithm, for simplicity reasons all results have been summarized in Table 1.

TABLE I. SUMMARY OF RESULTS:

Method:	***Linear Regression***	***K-Nearest Neighbor Regression***	***Random Forest Regression***	***AdaBoosting (Ensemble Method)***	***GradientBoosting (Ensemble Method)***
R2 Value:	.94	.91	.89	.66	.91
MSE Value:	122.856	217.0	250.454	669.221	192.176
RMSE Value:	11.084	14.731	15.826	25.869	13.863
MAPE Value:	679132480336900.375	1389110285064500.500	966413837924275.750	6179887029420057.000	932010946023252.625
OLS Value:	R2 - .95 Durbin-Watson – 1.977	R2 - .90 Durbin-Watson – 1.035	R2 - .90 Durbin-Watson - .884	N/A	N/A

As expected, all trained and tested regression models have performed extremely well. By setting the feature selection criteria to a .25 or higher correlation basis, the proposed research was successful in selecting key influencing features against the target variable. A clear runner up between the five undertaken regression methods is Linear Regression with an R2 value of .94, meaning that the proportion of variance of the dependent variable can be explained by 94% of the independent variables. However, by undertaking the Ordinary Least Regression analysis in conjunction to Linear Regression, KNN Regression and Random Forest. It was discovered that the data tested by Linear Regression is absent of autocorrelation which is the representation of degree of similarity as it measures the relationship between a variables current value and its past value. The Durbin-Watson tests this similarity, and it was discovered that the OLS result of Linear regression are 1.977 as illustrated by Table 1 which indicate a lack of autocorrelation.

One may consider this as a poor performance, thus looking to KNN regression or Random Forest regression as the next best performing models. However, with that comes an increase in all the other remaining evaluation methods such as the MSE and MAPE.

Lastly, the proposed research introduced two ensemble learning methods as an exploratory element with the hopes to benchmark against the initially selected regression methods. As boosting methods within ensemble learning combines weak learner into stronger learners to minimize training errors, it was hoped that be higher scores than the initial methods would be achieved. Based on the two selected ensemble boosting methods, AdaBoosting and GradientBoosting both performed adequately with the GradientBoosting method taking the lead and returning similar results to that of the Linear Regression model as per Table 1. Thus, in the end it has been determined that the best scoring method is of Linear Regression.

V. Conclusion and Future Work

The purpose of the proposed research was the application of simple yet robust machine learning methods primarily in a regression format to determine if weather statistics are effective in predicting flight delays as set out by the research question. Based on the summary of results from Table 1, the proposed research would further aspire to undertake future work where it would look to explore a mix of simple and deep learning machine learning methods. There are many additional factors such as airport volumes, security clearance and crew management which can influence a delay or even cancellation of a commercial flight. Thus, the proposed research hopes to grasp an understanding of deep learning methods with the aim of combining the researched with the various other factors which may have potential influence. Stacked Recurrent Neural Network models is a promising application for future work, where each layer can represent a different factor.

In conclusion based on the feature selection and train/test of the data prior to model application the proposed research can confidently deem that the selected simple, yet robust machine learning methods are just as effective in predicting flight delays as that of Deep Neural Networks. With Linear Regression achieving a remarkable score of 94%, the proposed research was successful in achieving result higher to the models utilising deep learning methods [4] where these methods have achieved on average 90% accuracy scores.

Acknowledgments

The research team would like to express their appreciation to National College of Ireland who has provided the team with the possibility to complete this study. A special thanks is given to the lecturer of Data Mining and Machine Learning 1, Barry Haycock.

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] Ioana Gheorghiu, "HistoricalFlight Delay and Weather data USA", 2020 Feb 1st, Accessed on Feb 23rd, 2022, [Online], Available at: https://www.kaggle.com/ioanagheorghiu/historical-flight-and-weather-data/version/1