Portfolio Project: Proposal / Interim Project

Data Mining and Machine Learning 1, MSc/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*—This document serves the purpose of an interim project paper outlining a project idea and an overview of machine learning methods the final project aims to apply to the chosen area. The proposal for the Portfolio Project holds no grading weight.**

Keywords—JFK: *John F Kennedy,* ***LAX:*** *Los Angelos International Airport,* ***NOAA:*** *National Oceanic and Atmospheric Administration,* ***NCEI:*** *National Centers for Environmental Information,* ***DNN:*** *Deep Neural Networks*

I. Motivation

The goal of this project is to explore and apply various machine learning methods to three relatively related datasets of large caliber. The project looks to select different variables from each dataset to evaluate the chosen variable to determine which of the selected are best at predicting delayed flights for the machine learning methods identified. The datasets supporting the outlined project contain variables of both binary and multivariate nature.

The motivation behind this work originates from the desire to understand the decisional factors undertaken to result in a delay or ultimately a cancellation to a commercial flight due to the weather forecast. According to the International Air Transport Association, air travel is the safest form of long-distance, with a fatal accident rate of 0.28 per million flights [1]. Thus, the curiosity to understand arises behind the weather factors which can influence a flight from departure to arrival.

The exploration and prediction of flight delays and or cancelation for numerous reasons are not a first in the spectrum of machine learning [2] however, for the identification of the most optimal predictors many continue the research in this field. The use of machine learning to predict flight delays or cancelation allows airlines and airports to stay efficient and optimize their revenues. The benefits derived from the said machine learning applications all depend on the accuracy of the predictive models thus great importance is placed on the build of the model and the accurate selection of the best predictor variables.

The project hopes to recommend for future studies predictors variables of optimal nature as well as machine learning methods that perform best for the different datasets at hand. It is noted that discrepancies may arise between the project proposal report and the final portfolio around the machine learning methods as mixed results can occur upon the application of methods to the chosen datasets.

II. Research Question

The proposed project aims to research and 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 project will look to define several objectives which the research question will attempt to meet.

III. Initial Literature Review

The application and use of machine learning for the prediction of flight delays is not a new practice and is widely used in today's world, all with a common goal of a level of optimization either for the airline, passengers, airports, or all three. It is to no surprise at the vast amounts of literature reviews available to the public for consumption. A common result since the inception of Artificial Learning and the use of deep learning which has become a dominant approach for much of the machine learning work [5].

A common application of supervised machine learning to the prediction of flights has been in the form of Deep Neural Networks with multiple research models applied and literature reviews released. The prediction of flight delays can be possible through several combinations such as Long Short-Term Memory and Recurrent Neural Network with models achieving accuracy levels between 86% to 87% [6]. Alternatively, various other combinations of DNN techniques have been applied to predict delays, all achieving high accuracy levels ranging from as low as 86% to 96% [7][8][9].

However, the interpretation of DNN algorithms is a complex activity and a certain level of expertise is needed for the consumption of DNN algorithm evaluations [10].

On the contrary to DNN algorithms, much simpler techniques can and have been applied to the exploration of predicting flight delays. Methods such as Support Vector Machines, Decision Trees, Random Forest, and Classification all with similarly high accuracy levels to DNN [9][11][12][13].

Previously published studies do not limit to the exploration of machine learning application of one technique at a given time. Approaches such as direct comparison between DNN algorithms and the much simpler methods looked to evaluate the comparison between Long Short-Term Memory and Random Forest [14]. Research on both machine learning methods yielded surprising results with the data worked, scoring high accuracy levels. Nevertheless, as with all methods and models, there are problems with both which

are taken into consideration and improvement steps considered.

Whilst there are vast amounts of published literature available, each operating and evaluating with different features and datasets. It has been noted that the successfulness of the mentioned works of literature is evaluated based on accuracy levels. The proposed portfolio looks to the identification of suitable machine learning methods in line with the datasets presents all whilst taking into consideration that the accuracy levels of prediction may be the driving force.

IV. Data Sources

The project is supported by three large datasets with a degree of similarity between all. The datasets being worked on are as follows:

1. The first dataset outlines all recorded flight delays for the United States of America for the year of 2019 in conjunction to weather descriptions on an hourly basis for a single record. This dataset is sourced from Kaggle [3].
2. The second dataset outlines the daily and hourly measurements of weather from the JFK Airport weather station for the period of the 1st of January 2020 to the 31st of January 2021. This dataset is sourced from the NCEI, NOAA [4].
3. The third dataset outlines the daily and hourly measurements of weather from the LAX Airport weather station for the period of the 1st of January 2020 to the 31st of January 2021. This dataset is sourced from the NCEI, NOAA [4].

V. Identification of Machine Learning Methods

The project looks to apply a total of five machine learning methods, with the approach to cover two methods as Unsupervised Learning and three as Supervised Learning.

The project hopes to uncover any hidden patterns within the data through the aid of Unsupervised Learning [15]. This is to be achieved through the application of the following methods.

1. Clustering is a common data mining technique that works by grouping unlabeled data based on either their differences or similarities. K-means clustering is to be employed to assist the grouping of the data points. The use of K-means clustering will allow for the identification of groups that are not explicitly labeled in the data. This machine learning method is to be used as an exploratory activity to better understand the underlining nature of the raw data.
2. Dimensionality Reduction is a data mining technique that simplifies the work involved with big data by reducing the number of data inputs into manageable sizes. The project looks to utilize Dimensionality Reduction to improve the performance of other machine learning methods by addressing overfitting, as well as simplifying the visualization of the big data at hand. The project looks to work with the Principal Component Analysis algorithm to reduce redundancies with the data and compress the datasets through feature extraction.

Furthermore, the project looks to the following Supervised Learning [16] methods to support the study in exploring and achieving high accuracy levels for future predictions.

3. Linear Regression, a popular technique both in Statistics and Machine Learning is used to identify the relationship between a y variable otherwise known as the dependent variable, and one or more x variables aka independent variables. As the study looks to explore the effects of weather on flight delays with the end goal of predicting future delays based on weather data, the project finds it suitable for a linear regression analysis to be evaluated.
4. K-NN Regression, compared to Linear Regression this method is non-parametric which states that the data at hand is not required to meet certain assumptions or parameters. The project looks to evaluate the comparison between the two regression techniques to identify the best-suited method for the data at hand. It is assumed that the intuitive manner behind KNN Regression can be just as equally effective as Linear Regression when the target variable is continuous.
5. Lastly, the project looks to explore the Random Forest machine learning method. This method is chosen alternatively to Decision Trees due to its ability to resolve overfitting. The method works by referencing a collection of uncorrelated decision trees which are then merged to reduce the variance to create more accurate predictions. Additionally, the project looks to improve the accuracy level through the application of Random Forest in place of the published works of literature employing the said method.

It is noted that the chosen machine learning methods for the proposed project may vary to the actual work as the final application may pivot the proposed approach to better suit the datasets at hand.

VI. Identification of Evaluation Methods

All machine learning algorithms utilized for the proposed project and their evaluation are to be undertaken through the programming language of Python. The environment to assist the programmatic structuring will be PyCharm.

The project looks to utilize the sklearn package found in Python as the core operator for the identified machine learning algorithms.

To aid the evaluation of the proposed Unsupervised Learning methods, the project looks to employ the use of various visualizations to plot the clustering results of K-Means, this is to be achieved using matplot in Python. We look to derive meaningful clusters from the K-Means method, this can be achieved through the evaluation property inertia, a key indicator that states the distance of the points within the cluster. We will be seeking clusters with the distance as low as possible between them.

Additionality, using Principal Components Analysis the study can determine the importance of certain features from the proposed datasets prior to the application of Supervised Learning methods. This is to be facilitated through multiple

visualizations which will assist in understanding the relationship between the x and y values rather than jumping straight into predicting the y values based on the x values.

Lastly, as the three Supervised Learning methods identified are of regression nature, the study will look to the evaluation of the R^2 value where we will look for the highest value possible based on the model presented. We look to support this evaluation with the aid of the Root Mean Squared Error, the Mean Absolute Percentage Error, and the Mean Squared Error.

Additionally, the study looks to undertake a Sensitivity Test which tests the identified methods by outlining a proportion of the actual positive cases which get predicted as positive or in other terms true positives.

REFERENCES

[1] International Air Transport Association, "Aviation Safety", Accessed on Feb 22nd, 2022, [Online], Available at: https://www.iata.org/en/youandiata/travelers/aviation-safety/

[2] Javier Herbas, "Using Machine Learning to Predict Flight Delays", 2020 Oct 17th, Accessed on Feb 22nd, 2022, [Online], Available at: https://medium.com/analytics-vidhya/using-machine-learning-to-predict-flight-delays-e8a50b0bb64c

[3] 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

[4] NOAA, "Climate Data Online", Accesses on Feb 23rd, 2022, [Online], Available at: https://www.ncdc.noaa.gov/cdo-web/

[5] Wikipedia, "Machine Learning", 2022 Feb 23rd, Accessed on Feb 24th, 2022, [Online], Available at: https://en.wikipedia.org/wiki/Machine_learning#Artificial_intelligence

[6] 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, Accessed on 25th Feb 2022, [Online], Available at: https://ieeexplore.ieee.org/abstract/document/7778092

[7] 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

[8] 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, Accessed on 25th Feb 2022, [Online], Available at: https://ieeexplore.ieee.org/document/9512525

[9] 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, Accessed on 25th Feb 2022, [Online], Available at: https://ieeexplore.ieee.org/abstract/document/9251206

[10] Jason Brownlee, Machine Learning Mastery, " A Gentle Introduction to Long Short-Term Memory Networks by the Experts", 2017 May 24th, Accessed on Feb 24th, 2022, [Online], Available at: https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/

[11] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), 2016, pp. 1-6, doi: 10.1109/DASC.2016.7777956, Accessed on 25th Feb 2022, [Online], Available at: https://ieeexplore.ieee.org/abstract/document/7777956

[12] Jun Chen and Meng Li. "Chained Predictions of Flight Delay Using Machine Learning," AIAA 2019-1661. AIAA Scitech 2019 Forum. January 2019, Accessed on 25th Feb 2022, [Online], Available at: https://arc.aiaa.org/doi/10.2514/6.2019-1661

[13] E. Esmaeilzadeh and S. Mokhtarimousavi, "Machine Learning Approach for Flight Departure Delay Prediction and Analysis", *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2674, no. 8, pp. 145-159, 2020, Accessed on 25th Feb 2022, [Online], Available at: https://journals.sagepub.com/doi/10.1177/0361198120930014

[14] 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, Accessed on 25th Feb 2022, [Online], Available at: https://ieeexplore.ieee.org/abstract/document/8903554#sec5

[15] IBM, "Unsupervised Learning", 2020 Sept 21st, Accessed on 26th Feb 2022, [Online], Available at: https://www.ibm.com/cloud/learn/unsupervised-learning

[16] IBM, "Supervised Learning", 2020 Aug 19th, Accessed on 27th Feb 2022, [Online], Available at: https://www.ibm.com/cloud/learn/supervised-learning