FLIGHT DELAY PREDICTIONS USING MACHINE LEARNING ALGORITHMS

ABHISHEK BIJU DAS

Abstract

Flight delays present a formidable challenge to aircraft operators, billion dollar airlines and their passengers alike. They impede daily life and have massive, often invisible financial, emotional and psychological impact on all parties involved. Thus the prediction of said delays and cancellations is invaluable information not only for damage limitation, but to study and grasp its underlying causes. This paper sees the use of multiple such algorithms such as Decision Trees, Neural Networks, Random Forest Regressions as well as Gradient Boosted Regressions to develop an accurate machine learning model for the same and to propose an optimized deep learning model study that can be undertaken in the future.

Chapter 1

Introduction

Flight delays have been studied and analyzed rigorously through extensive research in recent years propelled mainly by the spike in air travel demand. This increased air traffic has also led to an increase in flight delays due to the sheer volume and performance pressure the industry is consistently expected to maintain. Citing one particular such report by the FAA (Federal Aviation Administration), the aviation industry frequently is subject to monetary losses amounting to as much as **$32.9 billion** in **2010** alone (this figure accounts for US air traffic alone). The factors that map to this prevalent phenomenon are air traffic congestion due to increased passenger throughput, safety and maintenance issues causing delayed release of aircraft, adverse weather conditions that may impede safe operation of aircraft including critical procedures like takeoff and landing, as well as the propagation of delays through chained connected flight paths. The global threshold for a delayed flight is understood to be **15 minutes** past the expected time of arrival. This paper attempts to provide a functional model to alleviate through concentrated collaborative efforts the effects of flight delay propagation within their network.

1.1 MACHINE LEARNING

Machine learning is a field that has seen a technological boom in the last decade. The Artificial Intelligence revolution equips us with various tools powered by Machine Learning and Deep Learning that utilize various algorithms for constructing, training and predicting mathematical models using historical data. The AI boom fuels a wide variety of use cases from these models which may be of deep relevance to businesses attempting to streamline their operations. Image recognition, speech recognition, and natural language processing (NLP) are some examples of fields that can use Machine Learning models to derive critical insight.

The essential idea surrounding Machine Learning is the usage of mathematical algorithms to analyze historical data which is then piped into prediction models. When a hypothetical dataset with a missing property is supplied to this prediction model, for instance, the model will be able to determine (with a certain degree of accuracy) these target labels. Machine Learning models excel in establishing complex relationships that properties in large rosters of data possess, often invisible to the human eye.

1.1.1 Features of Machine Learning

1. Machine learning utilizes data to identify different patterns within a provided dataset.
2. By learning from previous data, it has the ability to automatically enhance its performance.
3. Machine learning is a technology that heavily relies on high-quality data for its functioning, as the predictions can only be as good as the data fed into the model.

1.1.2 Classification of Machine Learning Algorithms

1.1.2.1 Supervised Machine Learning

Supervised learning refers to a type of machine learning where machines are trained using labeled training data, enabling them to predict output based on that data. Labeled data consists of input information that has already been associated with the correct output.

In supervised learning, the training data acts as a supervisor, instructing the machines on how to accurately predict the output. This approach is similar to a student learning under the guidance of a teacher. The process of supervised learning involves providing both input data and corresponding correct output data to the machine learning model. The goal of a supervised learning algorithm is to discover a mapping function that establishes a relationship between the input variable (X) and the output variable (y).

1.1.2.2 Unsupervised Learning

Unsupervised Learning differs from its former counterpart in that it works through utilizing algorithms to cluster unlabeled datasets. They are capable of detecting the intricate and delicate relationships between these data groupings without human intervention for the actual learning.

The three main use cases of machine learning algorithms are clustering, association, and dimensionality reduction

1.2 MACHINE LEARNING ALGORITHMS

1.2.1 Logistic Regression Algorithm

Logistic Regression Algorithms utilize supervised learning in order to predict a categorical target variable when supplied with a set of independent variables. A common euphemistic example of the application of a logistic regression algorithm is to implement binary classification of a target image, for instance, a set of pictures of dogs, cats, oranges and apples could be classified as an animal or not using a logistic regression algorithm. The provided dependent variable must be categorical to use this algorithm, as it will classify a provided target as hit or a miss, there is no classification value in between.

Logistic regressions can be of significance as they have the capacity to classify continuous and discrete datasets.

1.2.2 Decision Tree Algorithm

The Decision Tree Machine algorithm is a broad spectrum of specialized algorithms that excels at classification problems, using the same fundamental methodology. The tree-structured classifier with internal nodes represent different features of the dataset, and branches of the tree are rules implemented by the algorithm to decipher at each step, what rule would give us the most accurate independent sets; this pattern is then replicated all the way down the tree.

The CART(Classification and Regression Tree) algorithm is the most widely used classification algorithm, which entails asking a simple yes or no question at each newly node and based on the answer split into further subtrees.

1.2.3 Random Forest Algorithm

The Random Forest machine learning algorithm is a supervised learning technique that works off of the sound principle of ensemble learning, which involves layering multiple classifiers to tackle a complicated dataset and improve the accuracy of the model.

It usually consists of multiple decision tree classifiers that work on different subsets of the dataset, and compiles results to obtain the best possible classification accuracy. Overfitting is a common problem encountered when using Random Forest decision trees, and can be solved by adding more trees.

The general outline of the algorithm is to create the random forest by compiling N-number of decision trees, upon which predictions are then created.

1.2.4 Neural Networks

The supervised Neural Network Machine Learning algorithm we will implement is the **Multi-Layer Perceptron (MLP)** that trains a function based on a training dataset. It specifically learns a nonlinear function approximator that utilizes hidden neuron layers that performs transformations (hyperparameter tuning) based on exhaustive computations. MLP's have an advantage in that they are capable of real-time (online learning) and are well-suited for Machine Learning deployment on the cloud.

Chapter 2

Project Proposal

The key ideas that facilitate the need to address what this paper attempts to do, is that the forecasted growth in air travel and the subsequent boom in aviation industry capitalization underline the significance it has to economies worldwide. Besides being able to transport persons and cargo across the world in a matter of hours, it greatly enhances the integration of an economy with the outside world. Furthermore, studying the various factors that contribute to on-time operations and passenger fulfillment and an unbiased investigation into factors that delay operations is necessitated.

Machine Learning provides us with a robust framework to study massive amounts of operations data and extract relationships between variables invisible to the human eye. The primary aim of this paper is to implement, compare and contrast various Machine Learning algorithms to predict to the highest possible degree of accuracy flight delay patterns.

Chapter 3

Resource Catalog

3.1 Requirement Analysis

Features of a proposed system and the resources required by the system to undertake the task that the system claims to successfully solve provides a mechanism to develop a comprehensive solution.

3.1.1 Python

Python is a versatile, high-level, and freely available programming language that is interpreted and supports both object-oriented and procedural programming paradigms.Python has gained popularity due to its clear syntax and readability. It is considered relatively easy to learn and is portable, allowing its code to be interpreted on various operating systems.

The language was created by Guido van Rossum, who was inspired by his favorite comedy group, Monty Python's Flying Circus. Python's source code is freely available, open for modification, and can be reused. Consequently, Python boasts a considerable number of users worldwide.

Chapter 4

Development Timeline

4.1 System Architecture

1. **Data Retrieval**
2. **Pre-Processing**
3. **Feature Extraction**
4. **Model Evaluation**

Data Retrieval

All the flight delay prediction models that are implemented and discussed in this paper use data collected and stored over a span of **3 years** from **2020-2023**. This data is curated and stored securely in SQL databases in the AWS cloud maintained by the company. Some examples of data columns that we work with are estimated departure time, scheduled departure time, takeoff time, taxi-in time before boarding, flight distance, taxi-out time after boarding, flight time (airborne time) etc. The unedited dataset comprises **129 columns and 96902 rows**. Specific data points have been redacted from this report to maintain institutional data security.

Cleaning

After obtaining the data in our python script, we have to clean this data of unwanted edge data points as well as incomplete data points comprising null values that may skew our model training. Traditional approaches dictate running the data through a classifier that performs some standard redundancy management. This minimizes indexing which in turn boosts algorithm

performance. Furthermore, techniques such as tokenization, stemming and lemmatization may be adopted to downscale derivationally associated context words.

Feature Extraction

After examining various sources from external research conducted on the subject, we have filtered out the parameters with formidable impact potential on accuracy of the model:

- Date of Departure
- Origin Airport
- Destination Airport
- Taxi-out time
- Actual Departure time
- Expected Departure time
- Flight Number

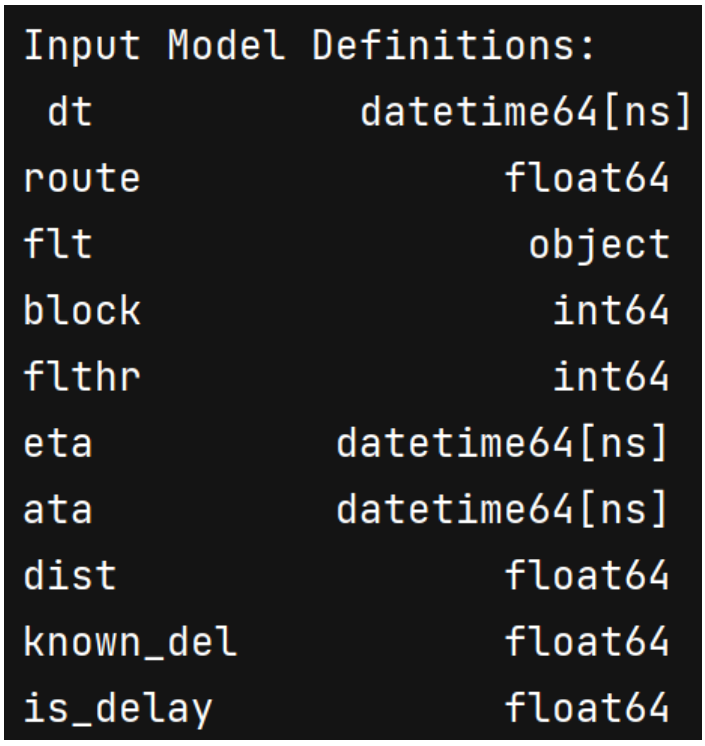

Figure 1. Model features that are used as input for all ML models (target label is is_delay)

Evaluation

Succeeding the cleaning and feature-extraction of the dataset, we create a **80-20** split for the training dataset and the testing dataset respectively. Accuracy metrics such as mean error percentage and accuracy of model is calculated with the powerful scikit-learn metrics package. For the purposes of this classification, we are going to consider any flight that touches down at its destination airport more than **15 minutes** past the expected touch down time to be a delayed flight. The model utilizes a binary prediction target label where a **0** stands for an on-time flight arrival at destination whereas a **1** stands for a delayed flight arrival at destination.

```
#creating the train and test data set
X_train, X_test, y_train, y_test = train_test_split(X,
    y, test_size=0.2, random_state= 123)
```

Figure 2. Generating the training and testing split for all models with a fixed random state

4.2 Machine Learning Algorithms

The different Machine Learning algorithms that have been explored in this paper's analysis are as follows. After unbiased consideration of all potential Machine Learning algorithms that boast relatively light computational costs and accurate results, shortlisted candidates were then tested to narrow down the most effective algorithm for the final model.

1. Linear Regression Models
2. Random Forest Decision Tree Models

3. Multi-Layer Perceptron Neural Network Classifiers
4. Gradient Boosted Decision Tree Models

The Linear Regression model on fitting and predicting provides us with predictions that are far too one-dimensional to extract enough viable insights from, and hence is disregarded for the rest of the paper. After much experimentation, feature extraction and training and testing, for the purposes of this paper, the best performance manufacturing Multi Layer Perceptron Neural Network Classifier has been chosen, as it seems to give us the largest model result accuracy of approximately **99.68%**.

dt	route	flt	block	flthr	eta	ata	dist	known_del	is_delay	predicted_is_delay
2014-12-31 00:00:00	583	583	738750	738750	1900-01-01 04:49:00	1900-01-01 04:42:00	3535	-7	0	1
2014-12-31 00:00:00	754	754	738750	738750	1900-01-01 03:34:00	1900-01-01 03:26:00	2965	-8	0	1
2014-12-31 00:00:00	331	332	738750	738750	1900-01-01 02:07:00	1900-01-01 01:56:00	1187	-11	0	0
2014-12-31 00:00:00	557	558	738750	738750	1900-01-01 01:39:00	1900-01-01 01:32:00	3285	-7	0	1
2014-12-31 00:00:00	590	589	738750	738750	1900-01-01 05:07:00	1900-01-01 04:56:00	3711	-11	0	1
2014-12-31 00:00:00	440	440	738750	738750	1900-01-01 06:27:00	1900-01-01 06:18:00	2943	-9	0	1
2014-12-31 00:00:00	631	632	738750	738750	1900-01-01 00:29:00	1900-01-01 00:21:00	2602	-8	0	1
2014-12-31 00:00:00	631	632	738750	738750	1900-01-01 00:29:00	1900-01-01 00:21:00	2602	-8	0	1
2014-12-31 00:00:00	155	156	738750	738750	1900-01-01 03:27:00	1900-01-01 03:22:00	2139	-5	0	1
2014-12-31 00:00:00	631	632	738750	738750	1900-01-01 00:29:00	1900-01-01 00:21:00	2602	-8	0	1
2014-12-31 00:00:00	707	707	738750	738750	1900-01-01 00:54:00	1900-01-01 00:47:00	1761	-7	0	1
2014-12-31 00:00:00	631	632	738750	738750	1900-01-01 00:29:00	1900-01-01 00:21:00	2602	-8	0	1
2014-12-31 00:00:00	41	41	738750	738750	1900-01-01 00:50:00	1900-01-01 00:41:00	348	-9	0	0
2014-12-31 00:00:00	557	558	738750	738750	1900-01-01 01:39:00	1900-01-01 01:32:00	3285	-7	0	1
2015-01-01 00:00:00	732	731	738750	738750	1900-01-01 05:39:00	1900-01-01 05:38:00	1443	-1	0	1
2015-01-01 00:00:00	551	551	738750	738750	1900-01-01 02:27:00	1900-01-01 02:18:00	169	-9	0	0
2015-01-01 00:00:00	3	4	738750	738750	1900-01-01 13:22:00	1900-01-01 13:23:00	382	1	1	1
2015-01-01 00:00:00	902	901	738750	738750	1900-01-01 10:31:00	1900-01-01 10:32:00	3526	1	1	1
2015-01-01 00:00:00	631	632	738750	738750	1900-01-01 00:31:00	1900-01-01 00:25:00	2602	-6	0	1
2015-01-01 00:00:00	1	1	738750	738750	1900-01-01 04:32:00	1900-01-01 04:32:00	382	0	0	1
2015-01-01 00:00:00	176	175	738750	738750	1900-01-01 23:11:00	1900-01-01 22:55:00	2587	-16	0	0
2015-01-01 00:00:00	44	44	738750	738750	1900-01-01 06:45:00	1900-01-01 06:37:00	348	-8	0	0

Figure 3. Snippet of predictions at a random index slice; predicted_is_delay is the algorithm's prediction column while the is_delay column is the target label.

Chapter 6

System Study

6.1 Feasibility Study

This project's feasibility is examined and proposed with a brief execution outline. The system has been streamlined to be an asset to the company and derives insights that build off of the company's main goal of, 'Taking more people, more places.'

6.1.1 Economic Feasibility

The economic outreach and impact the system generates is based on the accuracy of the predictions as well as the potential foresight the system possesses through which it can prepare to receive and act on procedure pipelines for unaccounted flight delays.

6.1.2 Technical Feasibility

The technical load imposed by the system, as well as the computer architecture and processing power required to host the model and run it at regular intervals to study its predictions constitutes the technical feasibility of the model. To deploy this model on the company's existing cloud infrastructure would be well within the company's scope and budget.

Chapter 7

Conclusion

The Machine Learning algorithms explored in this paper along with its implementation to predict flight delays provides significant insight into the scope for the utilization and deployment of such models in business bottlenecks, that can not be otherwise evaluated through intellect.

The Multi Layer Perceptron Neural Network Machine Learning model trains using the backpropagation algorithm on an array of **n*N** (n samples and N features) and an array of size n which contains the training sample target value. After fitting the model(training) we receive an output which can predict labels for the test data. The coefficient weights of the model are fine-tuned during the learning process. Below is the binary output produced by the model with a corresponding coefficient weight, performed with a standard testing and training split of **80-20**.

We observe that to calculate delay, the Random Forest Decision Tree model and the Gradient Boosted Decision Tree provide a reliable accuracy of **97.64%** and **96.84%** respectively, however the Multi-Layer Perceptron Neural Network delivers us a slightly higher accuracy of **99.68%**, and these are the minimum accuracy values found within these respective metrics. Random Forest Regression (from scikit-learn) gives us a mean squared error of **2215.4** and mean absolute error of **23.2** whereas the Gradient Boosted Decision Tree delivers us an mean squared accuracy of **2957.8** and a mean absolute error of **28.4**.

```
X_train.shape
 (76520, 1933)
X_test.shape
 (19131, 1933)
Accuracy Score for MLP Neural Network Classification Percentage: 99.68 %
```

```
Gradient Boosted Decision Tree Accuracy : 96.84%
Gradient Boosted Decision Tree Mean Squared Error(MSE) : 2957.8
Gradient Boosted Decision Tree Mean Absolute Error(MAE) : 28.4
```

```
Random Forest Decision Tree Accuracy : 97.64%
Random Forest Decision Tree Mean Squared Error(MSE) : 2215.4
Random Forest Decision Tree Mean Absolute Error(MAE) : 23.2
```

Figure 4. ML algorithm standard metric evaluation

The future scope and enforcement of this paper's findings may involve the utilization of more advanced feature-extraction, pre-processing and training techniques. There is a large scope for optimization established through improved random state analysis and enhanced sampling techniques as well as the upgrading of deployed deep-learning techniques. Furthermore the quality of the dataset can be improved with the addition of meteorological weather data variables that can, for instance, be scraped real-time from a reliable weather forecast API, for developing close to error-free models that can be deployed with confidence. Moreover, one can further upgrade model effectiveness if the individual records in the dataset can be chained so as to predict delay propagation through a flight network. This will even enable the model to categorically predict not only if a flight is delayed, but how many minutes it is delayed by with stunning accuracy.

There is also a possibility of achieving minute adjective metric control by layering different deep learning algorithms on top of each other, to extract compressed insights into supply chain bottlenecks that can be enhanced. Besides this there is also a possibility of deploying similar models at viable outposts, as the data from all these models can be further analyzed by yet another deep learning model, which can deliver us invaluable insight into delays across the globe and produce a real-time predictive model.

References

R. Balamurugan, A.V. Maria, G. Baranidaran, L. Mary Gladence and S. Revathy, "Error Calculation for Prediction of Flight Delays using Machine Learning Classifiers," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 1219-1225, doi: 10.1109/ICOEI53556.2022.9776709.

Khaksar, H., & Sheikholeslami, A. (2017). Airline delay prediction by machine learning algorithms. Scientia Iranica. DOI Citation

MLPClassifier from sklearn.neural_network. (open-source) MLPClassifier Documentation

Kazmi, Syeda M. (2022). Integrating NLP Techniques to Enhance Automated Essay Evaluation. Master's Thesis. DOI Citation

Brody, Ann. (2019) Flight delays cost $32.9 billion, passengers foot half the bill, Berkeley Business and Economics Article Link