Flight Delays Predictions using PySpark

Abed Al Kader Al Bakkour
School of Informatics
University of Skövde
Skövde, Sweden
a22abeal@student.his.se

***Abstract*— Flight delays are a critical issue that affects the aviation industry and its stakeholders. The analysis of flight data can help identify the factors that contribute to delays and provide insights into how to prevent them. In this project, the big data tool PySpark was utilized to analyze a dataset of 5+ million flight records for the year 2015. Three machine learning algorithms were applied: Logistic Regression, Gradient Boosted Trees, and Support Vector Machines to predict flight delays. The results demonstrate that the Gradient Boosted Trees algorithm performs the best with an accuracy of 67.27%, followed by Logistic Regression with 66.72%, and Support Vector Machines with 65.80%. These results demonstrate the potential of machine learning algorithms in predicting flight delays and provide valuable insights for airlines and other stakeholders to improve their operations and customer experience. The compute power of a cloud-based platform was leveraged for executing the algorithms, but the analysis was performed using PySpark.**

Keywords—Flight Delays, PySpark, Big Data Tool, Logistic Regression, Gradient Boosted Trees, Support Vector Machines, Pipelines, Binary Classification Evaluator, Supervised Machine Learning Algorithms.

I. INTRODUCTION

Flight delays are a common problem in the airline industry, causing inconvenience and frustration for passengers and airlines alike. According to the US Bureau of Transportation Statistics (BTS), over 20% of domestic flights in the US were delayed in 2019, causing significant inconvenience, financial losses, and safety concerns [1]. Delayed flights also impact the environment by increasing fuel consumption and greenhouse gas emissions [5]. Therefore, there is a growing interest in analyzing flight data to identify the causes of delays, predict and prevent them, and improve the overall performance and customer experience of the aviation industry.

This project aims to develop a machine learning model for predicting flight delays using a large dataset of 5 million records for the year 2015. The dataset contains various attributes, but due to the large size of the dataset and the high percentage of null values in some attributes, the most significant attributes that can help predict flight delays such as airline, origin airport, scheduled departure and arrival, departure delay, distance, and date (month, day, and day of the week) were selected, which could influence the flight's punctuality.

The main analysis problem/question in this project is to predict flight delays based on the selected attributes using machine learning algorithms and big data tool due to the large size of data set such as Apache Spark (PySpark) [4]. This can help airlines and other stakeholders to identify the factors that contribute to delays and take preventive measures to improve their operations, scheduling, and customer service. The challenges in this project include handling the large size of the dataset, dealing with missing values (data quality), processing time, storage, selecting the appropriate machine learning algorithms and evaluating the performance of the models.

The gain of solving the problem of predicting flight delays is multi-fold. Firstly, it can improve the efficiency and profitability of airlines by reducing the costs of delays and cancellations, optimizing flight schedules, and improving customer satisfaction and loyalty. Secondly, it can enhance the safety and reliability of air travel by identifying and addressing the causes of delays, such as weather conditions, technical issues, and airport congestion. Thirdly, it can reduce the environmental impact of aviation by minimizing fuel consumption and emissions caused by delays and diversions.

In the following sections of this report, a detailed description of the dataset and the pre-processing steps performed to clean and transform the data will be presented. The data engineering design will be presented as well. The machine learning algorithms used to predict flight delays and their performance will also be discussed. Finally, the insights and recommendations derived from the analysis and suggestions for future research directions will be provided.

A. Research questions/hypothesis

How can PySpark analyze large-scale flight delays data and develop accurate predictive models for flight delays.

II. BACKGROUND

Flight delays have been a persistent problem in the aviation industry, causing significant financial losses to both airlines and passengers [1]. To address this issue, various studies have been conducted to predict flight delays using machine learning techniques. One such technique is the use of PySpark, a distributed computing framework that can handle large-scale datasets. PySpark is widely used in big data processing, machine learning, and predictive analytics due to its ability to handle complex data processing tasks efficiently.

PySpark is a powerful tool for processing large-scale datasets in a distributed computing environment. It provides a unified programming interface for working with distributed data processing frameworks, including Hadoop Distributed File System (HDFS), Apache Cassandra, and Apache HBase [2]. PySpark is built on top of Apache Spark, an open-source big data processing engine that offers speed, scalability, and fault-tolerance.

The primary advantage of PySpark is its ability to handle large datasets that exceed the memory capacity of a single machine in addition to the ability of scaling horizontally by adding more nodes to the cluster [2]. Thus, PySpark's distributed architecture allows it to break down the data into smaller chunks and distribute them across a cluster of nodes, where each node processes a portion of the data in parallel.

This approach enables PySpark to process data much faster than traditional single-machine solutions.

In the context of flight delay prediction, PySpark can be used to preprocess and clean the data, train and evaluate supervised machine learning models, and generate predictions. PySpark's data processing capabilities can be utilized to filter and transform the raw data, removing irrelevant or missing attributes, and aggregating or summarizing the data to extract meaningful features [3]. PySpark can also be used to join multiple datasets and perform complex data transformations, such as pivoting, grouping, and sorting.

The use of PySpark in conjunction with supervised machine learning algorithms, such as logistic regression, gradient boosted trees, and support vector machines, has been shown to achieve higher accuracies compared to traditional machine learning techniques. These algorithms can learn from historical flight data and identify patterns and relationships between the features and the target variable [3], i.e., flight delays.

Overall, the use of PySpark in flight delay prediction has demonstrated promising results, providing valuable insights into the factors that contribute to flight delays and enabling airlines to take proactive measures to mitigate their impact.

III. METHODOLOGY

A. *Data*

The Flight Delays data is collected and published by the DOT's Bureau of Transportation Statistics, and it is an open-source data, it is provided on Kaggle and contains data on domestic flights occurring in the US for the year of 2015. The data includes information for each flight, such as: year, month, day, day of the week, airline, flight number, tail number, origin airport, destination airport, scheduled departure, departure time, departure delay, etc. The size of the dataset is 550+ MB with 5,819,079 records.

B. *Data Acquisition*

i. Libraries:

The PySpark library is a Python interface to Apache Spark, a distributed computing platform used for big data processing. 'SparkSession' is used to create a Spark session and functions module provides various data manipulation and aggregation functions. 'StringIndexer' and 'OneHotEncoder' are used for data encoding, while 'VectorAssembler' is used to combine multiple feature columns into a single vector column. 'LogisticRegression', 'GBTClassifier', and 'LinearSVC' are popular machine learning algorithms used for classification tasks. The 'BinaryClassificationEvaluator' is used to evaluate binary classification models. The 'matplotlib.pyplot' and pandas' libraries are commonly used for data visualization and manipulation tasks. The 'plotly.graph_objs' module provides interactive visualization capabilities, allowing for more complex and interactive plots. The 'prettytable' library is used to create formatted tables in Python.

ii. Loading Data:

- **SparkSession**: object is created with appropriate settings for the task. The config method is used to set the URL of the master node (local[*] means to use all available cores), and the parallelism parameter is set to 16, which determines the number of partitions to be used when processing data.
- **Schema**: is explicitly defined for the CSV file. By defining the schema, it can ensure that the data types are correctly inferred and provide efficient data processing, as PySpark can skip the schema inference step.
- **Read**: the file is read into a PySpark data frame by additionally setting 16 partitions after manipulating so many times, setting partitions can speed up processing when working with large datasets.
- **Shuffle**: after reading the CSV file, the data frame is shuffled using the 'orderBy' method, which helps to distribute the data uniformly across all partitions. Then, the 'persist' function is used to cache the data frame into memory and disk, which allows for faster access and improved performance when performing iterative operations on the data frame.

C. *Data Pre-Processing*

i. Data Exploration and Analysis:

a) In figure 1 the vertical stacked bar chart shows the delayed and non-delayed flights per airline. Results show that WN airline has the highest delayed flights.

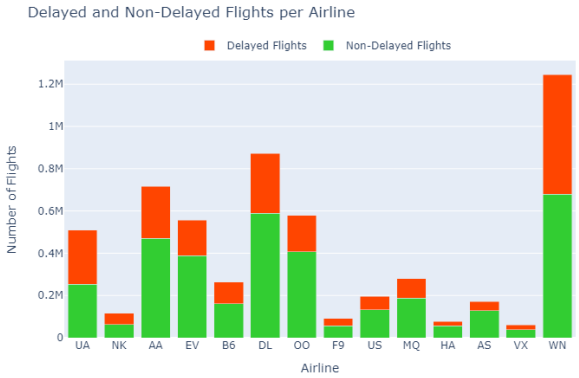

Figure 1

b) In figure 2 the pie chart shows the percentage of delayed and non-delayed flights. Results show that non-delayed flights are much higher than delayed flights with 62.9 % while delayed flights are 37.1% which means that there are unbalanced data.

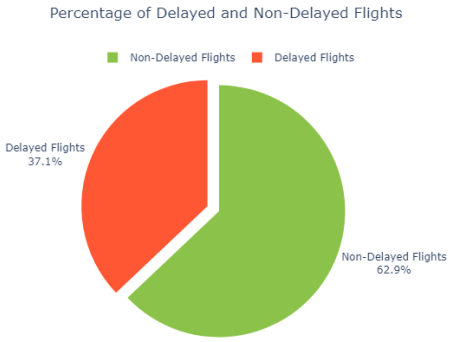

Figure 2

c) In figure 3 the bar chart shows the number of missing values. Results show that there are

extremely null values in many of the attributes. In this case, significant attributes will be selected in which it plays a big role in predicting delayed and non-delayed flights.

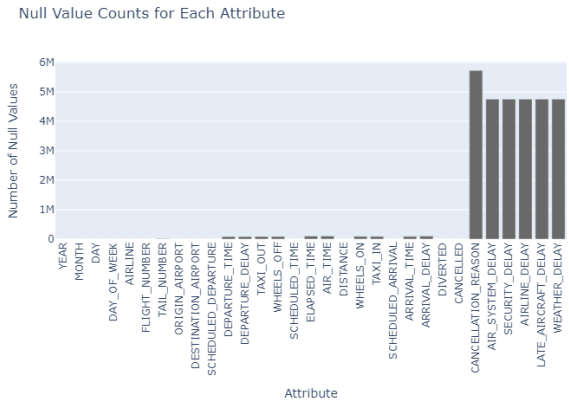

Figure 3

d) In figure 4 the scatter plot shows the distribution of flights over origin airports and month. The plot and results show that all flights that occurred in the month of October has incorrect IATA codes (airport codes), the strings are numbers. Thus, this means that 345K records should be dropped since these airport names can't be identified.

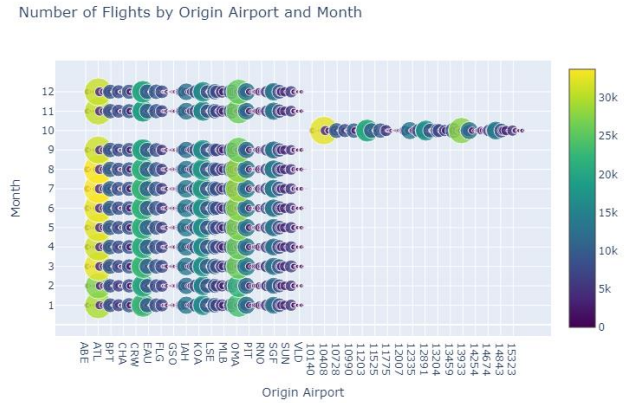

Figure 4

e) In figure 5 the line chart shows the frequency of number of flights with comparison to the number of delayed flights per month. Results show that the month of June has the highest number of delayed flights.

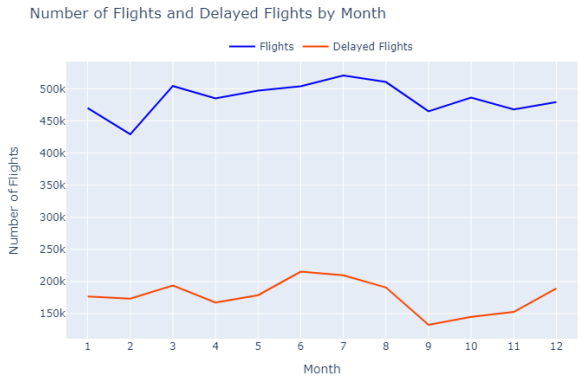

Figure 5

ii. Data Cleaning:

Significant attributes that can help in predicting whether a flight is delayed or not are selected such as month, day, day of week, airline, origin airport, scheduled departure, scheduled time, distance, scheduled arrival and departure delay. Then null values are cleaned by dropping them. In this case dropping 4K to 5K records will not affect the work as long as the data is still in millions, and it will improve the work of PySpark. Afterwards, the memory and disk are freed up since the attributes are assigned to a new data frame using the 'unpersist' function. The new aggregated data frame is then cached to the memory and disk for a faster access using the 'persist' function. Finally, the duplicates are checked and removed noticing that only 38 duplicates were found.

iii. Data Preparation

A new attribute was assigned to the dataset and named as 'Delay'. This attribute informs whether a flight is delayed or not by assigning 0 value for non-delayed flights and 1 value for delayed flights. This process facilitates the data engineering part which will be discussed later on. Previously, in the data exploration and analysis section the pie chart visualized and informed that the data is imbalanced, and this can lead to a model that is biased towards the majority class which is non-delayed flights and performs poorly on the minority class which is delayed flights. In this case it is required to balance the data, thus, under-sampling the majority class was one of the choices. So, after under-sampling still the data holds around 74% of the original data which is still perfect. Freeing up the memory and disk was applied, then cashing the new data frame which holds the new balanced dataset was tackled. Again previously, in the data exploration and analysis section the scatter plot showed that there are flights that occurred in the month of October holding wrong IATA codes as string of numbers. In this case, removing these records is also required. October month includes 344K records. It is still reasonable to remove these records because still the data is around 4.3M.

D. Feature Engineering

One hot encoding technique using PySpark is applied on attributes such as month, day, day of the week, airline and origin airport. This technique converted these categorical variables into numerical variables then used them as input features for the machine learning models. The encoders went afterwards into a pipeline as stages. The output after applying this technique is a sparse vector and is used as input to the vector assembler that has been created afterwards. Freeing up the disk and memory is required in this case by using the 'unpersist' function, then the vector assembler combined all features into a single vector which was the input of the machine learning models. Finally, the dataset is splitted into training and testing by taking 80% of the data for training the models and 20 % of the data for testing the models, then they are cached into memory and disk for faster access.

IV. MODELS

Three supervised machine learning algorithms were constructed using PySpark in this project to predict delay flights.

A. Logistic Regression

Logistic regression function is created using PySpark to create a logistic regression model by passing the parameter features that is created by the one hot encoder then passed to the vector assembler, and the parameter labels which is the 'Delay' that was created in the pre-processing stage. The 'Pipeline' function is then used to define the pipeline, which is a sequence of stages. In this task, the pipeline has only one stage, which is the logistic regression model. The advantages of using pipeline in this case are that it makes it easy to automate data pre-processing steps, such as scaling, encoding, and imputation. It also helps avoid The fit function is used then to fit the model on the training data. Then the transform function is used to make predictions on the test data.

B. Gradient Boosted Trees

Gradient Boosted Trees function is created using PySpark to create a gradient boosted trees model by passing the same parameters as logistic regression and same stages is applied such as pipelining, fitting the model onto training and predicting using the transform function. The GBT algorithm has the advantages of being able to handle large and discrete data sets, being applicable to both regression and classification issues, and being able to pinpoint the most crucial features in the data set.

C. Support Vector Machines

Support Vector Machines function 'LinearSVC' is created using PySpark to create an SVM model by passing the parameters maxIter of value 10, regParam of value 0.1, features and label which is 'Delay'. Then same stages such as pipelining, fitting and predicting is applied. The maxIter parameter of the LinearSVC model specifies the number of iterations that can be used to perform the optimization method. The optimization algorithm is used to find the ideal model parameter values that minimize the loss function. The regularization parameter, known as regParam, controls how much regularization is applied to the model. Regularization prevents overfitting by including a penalty term in the loss function.

V. EVALUATION

PySpark machine learning class such as the binary classification evaluator is used to evaluate the performance of the binary classification models LR, GBT and SVM by passing the parameters label which is 'Delay' and raw prediction column which is of type vector n this particular case which is rawPrediction. Afterwards, several evaluation metrics is computed by the binary classification evaluator including the accuracy, Area Under the Receiver Operating Characteristic Curve (ROC AUC) and the Area Under the Precision-Recall Curve (PR AUC).

VI. RESULTS

The evaluation of the models using binary classification evaluator showed that Gradient Boosted Trees had the highest accuracy of 67.27%, while Support Vector Machines had the lowest accuracy of 65.80%. The Logistic Regression had an accuracy of 66.72%. The metrics used for evaluation included accuracy, ROC AUC, and PR AUC, and the results were presented in figure 6.

Model	Accuracy	ROC AUC	PR AUC
Logistic Regression	66.72%	66.72%	66.16%
Gradient Boosted Trees	67.27%	67.70%	67.27%
Support Vector Machines	65.80%	66.54%	65.80%

Figure 6

VII. DISCUSSION

The results of the project indicate that the three machine learning algorithms used in this study, namely Logistic Regression, Gradient Boosted Trees, and Support Vector Machines, all perform relatively similarly in predicting flight delays. This suggests that all three algorithms could potentially be useful in predicting flight delays, and the choice of algorithm may depend on other factors such as interpretability, scalability, and computational resources. The results are significant to various stakeholders in the airline industry, including airline companies, airport authorities, and passengers. Accurate prediction of flight delays can help airlines to better manage their resources, such as aircraft and crew, reduce costs associated with delayed flights, and improve customer satisfaction. Airport authorities can use this information to optimize airport operations and minimize congestion. Passengers can benefit from timely and accurate information about flight delays, which can help them plan their travel itineraries and reduce the inconvenience associated with delays.

A. Limitations and Challenges

The results of this project suggest that handling large datasets in PySpark can present several challenges such as processing power and memory limitations can lead to long processing times and potential memory overflow errors. In addition to data preparation, cleaning, and pre- processing can be time-consuming and require careful consideration of the most important attributes to be used in the analysis. Moreover, visualizing the data can be challenging, as traditional data visualization methods may not be appropriate. Potential issues with data quality, such as missing or incomplete data, can affect the accuracy of the models. As seen in figure 3, there were attributes that could be significant in predicting flight delays such as weather delays, but due to the extreme missing values, there were no other choice than dropping these columns. Last but not least, working with such a large dataset may require specialized skills and expertise to ensure efficient processing and accurate results.

VIII. CONCLUSION

In conclusion, this project has provided valuable experience in using PySpark to handle and analyze large-scale flight delays data. The research question, "How can PySpark analyze large-scale flight delays data and develop accurate predictive models for flight delays?" led to the development of various hypotheses, which were tested through data

collection, management, analysis, visualization, and interaction. The findings showed that PySpark can efficiently handle 5 million flight delay records, and through the use of machine learning algorithms, accurate predictive models for flight delays can be developed.

The data management process involved cleaning and pre-processing the data, which resulted in the removal of some attributes with a large number of null values. The data analysis and visualization phase involved the use of various PySpark libraries to analyze the data, identify trends and patterns, and create visualizations for better understanding. Machine learning algorithms such as Logistic Regression, Gradient Boosted Trees, and Support Vector Machines were used to develop accurate predictive models for flight delays. The results showed that Gradient Boosted Trees performed the best in terms of accuracy, ROC AUC, and PR AUC.

However, the project has some limitations. Thus, to improve the accuracy of the predictive models, additional investigations need to be performed, such as incorporating real-time data, adding more features to the dataset, and testing different machine learning algorithms. Overall, the project provided a great opportunity to gain practical skills in handling and analyzing big data using PySpark and machine learning algorithms to develop accurate predictive models for flight delays.

REFERENCES

[1] J. Chen and M. Li, "Chained Predictions of Flight Delay Using Machine Learning," 2019.

[2] T. Nibareke and J. Laassiri, "Using Big Data-machine learning models for diabetes prediction and flight delays analytics," *Journal of Big Data*, vol. 7, p. 78, 2020.

[3] R. Sausen and U. Schumann, "Estimates of the Climate Response to Aircraft CO2 and NOxEmissions Scenarios," *Climatic Change*, vol. 44, p. 27–58, 2000.

[4] T. Drabas and D. Lee, Learning PySpark, Packt Publishing Ltd, 2017.

[5] R. Khan, S. Akbar and T. A. Zahed, "Flight Delay Prediction Based on Gradient Boosting Ensemble Techniques," pp. 1-5, December 2022.