Department of Decision Science

Faculty of Business

University of Moratuwa

Semester 08

DA4120 - Business Intelligence

Group Assignment: Analyze Big Data for creating Business Value

Department of Decision Science

Faculty of Business

University of Moratuwa

Semester 08

DA4120 - Business Intelligence

Group Formation Sheet

Group 05

No	Index number	Name
1	206113B	S.P.C.H. SENADEERA
2	206124J	S.S.N. SIRIMANNA
3	206021P	P.A.S.M. BANDARA
4	206089E	K.P.B.C.M. PATHIRANA
5	206150J	Y.I. WANIGATUNGA

TABLE OF CONTENTS

INTRODUCTION

In this assignment, there is a main challenge of carrying out data analysis on large dataset for the purpose of providing insights to a young hypothetical business. For this project, we have selected a Logistics and Cargo company, an important industry where the application of big data analysis can bring a lot of improvement to the services offered and, therefore, the company's revenues. With the help of the tools and approaches of data analysis and data mining, we are going to achieve the goal of making effective logistics decisions at the company, as well as foreseeing the potential trends and making rational decisions on the distribution of resources among them.

The above picture demonstrates the following steps we have taken: Firstly, we acquired the dataset from Kaggle; secondly, we conducted Exploratory Data Analysis (EDA) to analyze the dataset and discover the underlying patterns. Data cleansing is done to the dataset and then uploaded on Azure Blob Storage for secure access anytime because of its scalability. It is then possible to use Azure Analysis Services to come up with an even better model of data handling, which is well structured and presents a lot of meaning. Microsoft Visual Studio is used to construct an Analysis Services Tabular Project of which we extract the dataset from Blob Storage while creating a consolidated and stable data model.

This is followed by deploying the data model to Azure Analysis Services where it establishes a live connection to Power BI where our data is further analyzed. This approach makes it possible to perform real-time and large scale analytics and get insights about the logistics and cargo business. The last is to use the processed data in Power BI to make business conclusions of trends, improve areas that need optimization and make actionable business conclusions for business success.

By so doing, it is our desire to offer the structured data analysis that will be essential in helping the Logistics and Cargo company to make better decisions hence improving its operations as a way of growing the company in a competitive environment.

DATA DESCRIPTION

Dataset used: Airline Delay and Cancellation Data, 2009 – 2018

Dataset link: Airline Delay and Cancellation Data, 2009 - 2018 (kaggle.com)

The data set selected for this analysis is 'Airline Delay and Cancellation Data (2009-2018)' get from Kaggle. For this analysis, only the data of the year 2016, 2017 and 2018 has been utilized and all the three files in their combined form. This dataset contains the records of airline delays and cancellations in different airports in USA and it is designed to provide year over year analysis of airline performance. The information has been obtained from the U. S. Department of Transportation Bureau of Transportation Statistics (BTS) that monitors flight operations in the United States.

Containing close to 20 million records this dataset can be described as very large, making it suitable for big data analysis. The use of this dataset would allow achieving the goal of deriving useful business insights of airlines that would help airline companies or airports in achieving operational efficiencies, minimizing delays, and increasing customer satisfaction. In addition, the analysis will look at the causes of flight delay and cancellation; and applying statistical and machine learning models to forecast possible delays. This will in the end benefit the airline businesses in enhancing data decisions in a bid to reduce operational cost and at the same time enhance passengers' experiences.

Data pre-processing, cleaning, exploration and analysis data visualization will be a key part of this analysis. The outcomes will be mean to be used in generating information that can add value to airlines or businesses related to it. The codes, scripts, and configurations used in this analysis will be made available for replication, moreover, this document brings out the methodologies involved, problems faced, and the overall business value derived from the data analysis for the users.

DATA PREPROCESSING

Loading the dataset

This is a big data related procedure. So the dataset size is very large. Hence we cannot use standard data processing techniques in order to load the dataset. To handle this, we used PySpark, which is a Python API for Apache Spark. PySpark is designed for big data and allows for parallel processing, which makes it an ideal tool for managing large datasets. Below steps are used to load the dataset:

- Mounting Google Drive

```
from google.colab import drive
drive.mount('/content/drive')
```

In order to access data saved on Google Drive, this step mounts Google Drive to Colab. In this case, a specific directory (/content/drive/MyDrive/BI_Assignment/) contains the dataset (combined_output.csv).

- Initializing PySpark

```
from pyspark.sql import SparkSession
import pandas as pd
```

The purpose of importing these libraries is to start a Spark session, which is required to create a connection with Spark. “Pandas” may still be used as a bridge when working with smaller datasets or for smaller jobs, even if PySpark is designed to handle huge data processing.

```
# Initialize Spark Session
spark = SparkSession.builder.appName("Airline Delays Analysis").getOrCreate()
```

- Creating a Spark Session

When a Spark session is started, it is given the application name "Airline Delays Analysis." The “getOrCreate()” function makes sure that a Spark session is reused or that a new one is created.

- Loading the Dataset

```
df= spark.read.csv('/content/drive/MyDrive/BI_Assignment/combined_output.csv', header=True, inferSchema=True)
```

Next, the “read.csv” method from PySpark's “DataFrame” API is used to read the dataset. While “inferSchema=True” enables Spark to automatically identify the data types for each column, the “header=True” option guarantees that the first row is utilized as the column names.

Analysis of missing data

The dataset contains airline delay and cancellation records from 2016 to 2018. Here the main goal is to figure out the existence and amount of missing data in this dataset. This is an important phase since it impacts directly on how strong the analytical and decision-making procedures that follow are.

In order to overcome the problem of missing data, a PySpark script was created and run in order to determine the number and percentage of missing values in each of the dataset's columns. With the dataset being over a gigabyte in size, PySpark was selected because of its credibility for handling huge datasets with efficiency using distributed computing.

Column	MissingCount	MissingPercentage
FL_DATE	0	0.0
OP_CARRIER	0	0.0
OP_CARRIER_FL_NUM	0	0.0
ORIGIN	0	0.0
DEST	0	0.0
CRS_DEP_TIME	0	0.0
DEP_TIME	256081	1.3837933936660143
DEP_DELAY	261033	1.4105526803191986
TAXI_OUT	263393	1.4233054905981797
WHEELS_OFF	263388	1.4232784719323344
WHEELS_ON	271764	1.4685401409563796
TAXI_IN	271764	1.4685401409563796
CRS_ARR_TIME	0	0.0
ARR_TIME	271763	1.4685347372232105
ARR_DELAY	311764	1.6846894677187734
CANCELLED	0	0.0
CANCELLATION_CODE	18240587	98.5672649950218
DIVERTED	0	0.0
CRS_ELAPSED_TIME	23	1.2428586288837644E-4
ACTUAL_ELAPSED_TIME	309166	1.6706505689455562
AIR_TIME	309166	1.6706505689455562
DISTANCE	0	0.0
CARRIER_DELAY	15159303	81.91682844092841
WEATHER_DELAY	15159303	81.91682844092841
NAS_DELAY	15159303	81.91682844092841
SECURITY_DELAY	15159303	81.91682844092841
LATE_AIRCRAFT_DELAY	15159303	81.91682844092841
Unnamed: 27	18505725	100.0

According the figure, the analysis showed a varied distribution of missing data across different columns. Significantly, some columns such as CANCELLATION_CODE and specific delay reasons like CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, and SECURITY_DELAY displayed very

high missing data percentages, frequently exceeding 80%. This pattern suggests that these events are infrequent or only recorded under certain conditions. In contrast, essential flight details like FL_DATE, OP_CARRIER, and ORIGIN showed complete data availability, underscoring their consistent reporting across all entries.

After identifying the extent of missing data in the dataset, strategic methods were implemented to cleanse the dataset. This process involved selectively dropping columns and rows with excessive missing values and imputing others where reasonable assumptions could be made.

```
from pyspark.sql.functions import col

# Step 1: Drop columns with more than 90% missing values
# You can manually drop based on inspection, or automate based on missing percentage.
columns_to_drop = ['CANCELLATION_CODE', 'Unnamed: 27']
cleaned_df_spark = df.drop(*columns_to_drop)

# Step 2: Fill missing values for delay-related columns with 0 (since a missing delay likely indicates no delay)
columns_to_fill_zero = ['DEP_DELAY', 'TAXI_OUT', 'WHEELS_OFF', 'WHEELS_ON', 'TAXI_IN',
                        'ARR_TIME', 'ARR_DELAY', 'CARRIER_DELAY', 'WEATHER_DELAY',
                        'NAS_DELAY', 'SECURITY_DELAY', 'LATE_AIRCRAFT_DELAY']

# Filling missing values with 0
cleaned_df_spark = cleaned_df_spark.fillna(0, subset=columns_to_fill_zero)

# Step 3: Drop remaining rows that still have missing values in critical columns like flight times
critical_columns = ['DEP_TIME', 'ARR_TIME', 'ACTUAL_ELAPSED_TIME']
cleaned_df_spark = cleaned_df_spark.dropna(subset=critical_columns)

# Step 4: Display the cleaned DataFrame to verify
cleaned_df_spark.show(5)
```

Here the columns CANCELLATION_CODE and Unnamed: 27 were removed due to their high percentage of missing data (more than 90%). This decision was based on the insight that retaining these would not contribute significantly to further analyses due to the lack of comprehensive data.

Also, for columns related to delays (DEP_DELAY, ARR_DELAY, etc.), missing values were filled with zero. This approach was adopted under the assumption that the absence of recorded delay data corresponds to no actual delay, thus zero is a logical substitute.

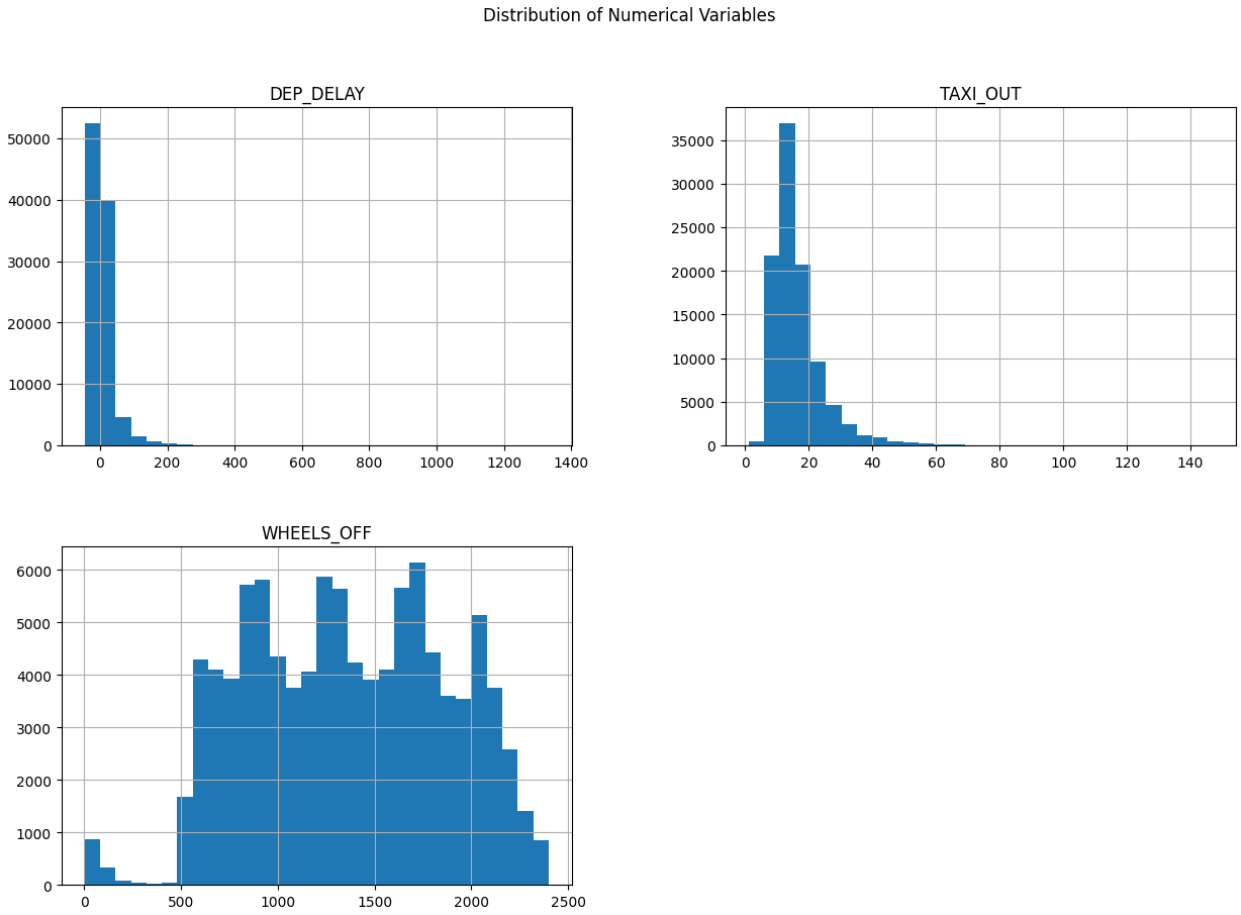
Dropping rows is also one remedy that we used. Rows, missing critical flight time information (DEP_TIME, ARR_TIME, and ACTUAL_ELAPSED_TIME) were dropped. The entire data for these parameters is required since they are critical to any meaningful analysis of flight performance and delay trends.

After using all these techniques, the dataset was much improved, providing a more stable basis for carrying out in-depth investigations of aircraft delays and cancellations. After being cleaned, the data significantly reduced any biases or errors that may have resulted from untreated missing data, making it ready for further in-depth exploration and predictive analytics.

EXPLORATORY DATA ANALYSIS (EDA)

The major aim of the exploratory data analysis is to get a feel of the distribution of data and identify exceptional features if any or important trends in data. This stage is the foundation to many analytical or predictive modeling exercises; it enables the identification of the relationships between the variables as well as the properties of the data.

Histogram Analysis



Under the EDA part we first create 3 histograms. In frequency distributions, graphical methods used are histograms, which helps one identify how frequently a value occurs in the several numerical characteristics in the dataset. From the histograms generated from the DEP_DELAY, TAXI_OUT and WHEELS_OFF columns, one can gain some average operating values of the airlines.

If we focus about DEP_DELAY (Departure Delay):

- This histogram shows a highly skewed distribution, with the majority of flights having little to no delay. The peak is very close to zero, suggesting that most flights depart on time or close to their scheduled departure. However, there are long tails extending to the right, indicating

occasional significant delays. These extreme values could be due to various factors such as weather conditions, mechanical issues, or air traffic control constraints.

The next histogram is about TAXI_OUT:

- The 'Taxi Out' time is defined as the duration from gate departure until takeoff. The histogram shows a somewhat right-skewed distribution but less extreme than the departure delays. Most taxi-out times cluster between 10 to 20 minutes, which aligns with typical airport operations. There are fewer flights experiencing very long taxi times, which are often influenced by airport congestion or operational inefficiencies.

The third one is WHEELS_OFF:

- The 'Wheels Off' time marks the point when the aircraft's wheels leave the runway. The distribution displayed across multiple peaks may indicate different waves of flight departures typical at large airports, reflecting time-blocked takeoff schedules. The spread and multiple peaks in this histogram suggest that departures are staggered throughout the day, correlating with airport slot times and peak traffic periods.

These histograms are very helpful to study the management aspects related to flight schedules and may look for ways to optimize and minimize the general delay time. Navigating through these variables enables identification of the operational surmises at airports and improved means at the airport.

Correlation Matrix

After performing an examination of each of the numerical feature separately, we proceeded with the examination of the interrelation between these variables using a correlation matrix. This allows in determining the direction and the degree of correlation between the variables and this plays a key role for instance in evaluating areas that may cause delay among airline operations.

The correlation matrix includes three key operational metrics: DEP_DELAY, TAXI_OUT and WHEELS_OFF. Here is what the matrix reveals:Here is what the matrix reveals:

- DEP_DELAY and TAXI_OUT (Correlation: It has been seen that the community has a higher level of education and as a consequence, there is a higher literacy rate: (0. 04).

Weak positive relationship between departure delay and taxi-out time are interpreted as there is a slighth positive linearity between them. This means that, taxi-out times could be related to minority departing delays with no very close or significant connection as to suggest that taxi-out has a direct influence to it.

- DEP_DELAY and WHEELS_OFF (Correlation: Therefore (or ‘thus’, 0. 13):

The two factors which have a somewhat related nature of the wheels leaving the runway and the aircraft departure delays and that there is a quite a low correlation between the two variables. This correlation has implications that mean flight that departs later than the set time usually has experienced a level of delay at the time of take-off. While there is a positive relationship between the two, the results suggest that it is quite moderate meaning that there are other factors that strongly determine departures delays.

- TAXI_OUT and WHEELS_OFF (Correlation: Thus, the aimed values for the innovative product and its price are the following: pA = 400 ; pB = 400 − 0. 027*400 = 400 − 10. 8 = 389. 2.

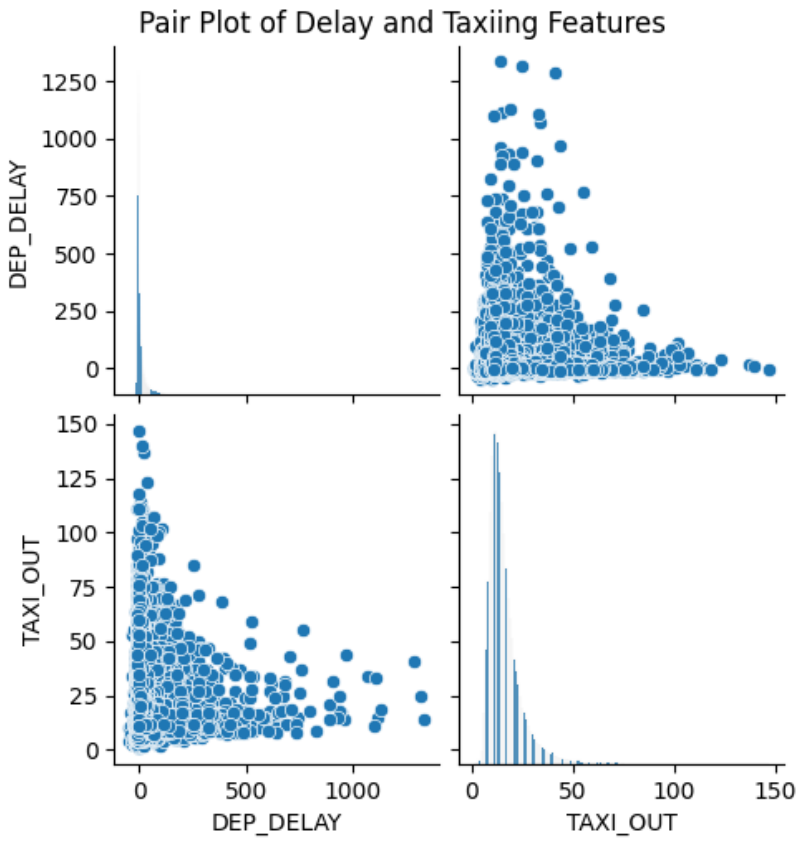
It is a negative correlation – very weak and almost insignificant which means there is really no strong, or any for that matter, linear association between the taxi-out time and the wheels-off time. This means that changes in taxi-out time do not correspondingly translate to changes in the number of minutes it takes for the wheels of the aircraft to rise off the ground.

By analyzing the above correlation matrix it is possible to establish that the underlying correlations between these variables are rather weak, meaning that as much as one may be able to see some form of linear relationships existing between the variables under analysis, it is not very strongly defined in such a way that we can conclusively say that how a given variable behaves depends on another variable as demonstrated below. This may be a pointer to the fact that these operational measures are subject to the impact of several other external factors including the flow of business through airports, ATC decisions, and weather patterns among others that all have substantial roles to play in the determination of these measures.

Although useful in identifying potential correlation, this matrix should be followed by a more detailed statistical analysis including possibly development of predictive models to better understand relations occurring in the airline environment.

Analysis of the pair-plot

In this section the pair plot provides a detailed visual comparison between DEP_DELAY and TAXI_OUT through scatter plots and histograms. This type of visualization helps in understanding both the distribution of individual variables and the relationships between them.



Description of Plots:

- Top-Left (DEP_DELAY vs. TAXI_OUT):

This scatter plot demonstrates that data points are spread out randomly, while having a focus on the area where both variables, the departure delays, and taxi-out time, are low, which corresponds to normal activity. The taxi out times slightly increase with departure delay regardless the delay type, although a sharp increase is observed in delays up to approximately 200 minutes for air carrier flights. After this the relationship is not quite as distinct which may be due to there being less points of comparison in patients with very long wait times.

- Top-Right and Bottom-Left (Histograms):

These histograms provide the distribution of DEP_DELAY and TAXI_OUT, respectively. The DEP_DELAY histogram indicates a heavily right-skewed distribution, where most flights experience little to no delay. In contrast, the TAXI_OUT histogram shows a more spread out distribution, but still right-skewed, with most taxi-out times falling within a relatively short range.

- Bottom-Right (TAXI_OUT vs. DEP_DELAY):

This is the same as the top-left plot but rotated, emphasizing the same relationships from another perspective.

We can see that as relationship insights, there is a significant positive correlation between departure delays and taxi-out times, although it is not particularly strong. Increased taxi-out times could be contributing marginally to departure delays, or both could be influenced by common underlying factors such as airport congestion.

Understanding these relationships can help airports and airlines optimize their scheduling and ground operations to minimize taxi times, which could potentially reduce overall delays. Also, airline and airport operations teams can use this data to identify specific thresholds where increases in taxi-out times significantly impact departure delays and target interventions accordingly.

The analysis of these plots underlines the complexity of airline operations where multiple factors contribute to delays. By identifying and understanding patterns such as those revealed in these pair plots, operational strategies can be better formulated to enhance efficiency and reduce passenger inconvenience.

ML MODEL

Airline delays have significant consequences for airlines, passengers, and the broader transportation network. To address this issue, predictive models can offer valuable insights into delay patterns, enabling better decision-making and resource allocation. This project focuses on predicting airline delays using two machine learning models: **Logistic Regression** and **Random Forest Regression**, implemented using **PySpark**, a distributed computing framework. These models allow us to classify flights as delayed or not delayed and predict the actual delay times.

Tools and Technologies

For this project, several tools and technologies were used to build, train, and evaluate the model:

- PySpark: A scalable data processing framework used for handling large datasets and performing machine learning tasks. PySpark was chosen for its ability to efficiently process big data across distributed systems.
- LogisticRegression (MLlib): A widely used classification algorithm suitable for binary classification tasks. Logistic Regression models the probability of a binary outcome, in this case, whether a flight was delayed (1) or not (0), based on one or more predictor variables.
- Random **Forest Regression**: An ensemble learning method used to predict continuous variables such as delay times. Random Forests aggregate the predictions from multiple decision trees, making the model more robust to overfitting.
- MulticlassClassificationEvaluator: Used to evaluate the performance of the Logistic Regression model by calculating the accuracy of predictions.

- Root **Mean Squared Error (RMSE):** The metric used to evaluate the accuracy of the Random Forest Regression model in predicting delay times.

Classification Model Selection

Logistic Regression

We used the Logistic Regression model for predicting whether a flight will be delayed or not since it is simple algorithm for binary classification. Logistic Regression is a linear model which predicts the probability of a binary dependent variable based on the independent variables or features. It is particularly useful for interpretable models where we need to understand a relationship between features and prediction outcome.

Random Forest Regression

To estimate the actual delay times we applied regression using the Random Forest method. This is an example of the voting model where the construction of many decision trees takes place and the results are then averaged in a bid to try and make it less volatile and more generalized. Continued variable prediction is the best suited for the Random Forest Regression algorithm, and the best suited to manage contextually intricate relationships within the data set and also control the possibility of overfitting.

Model Training

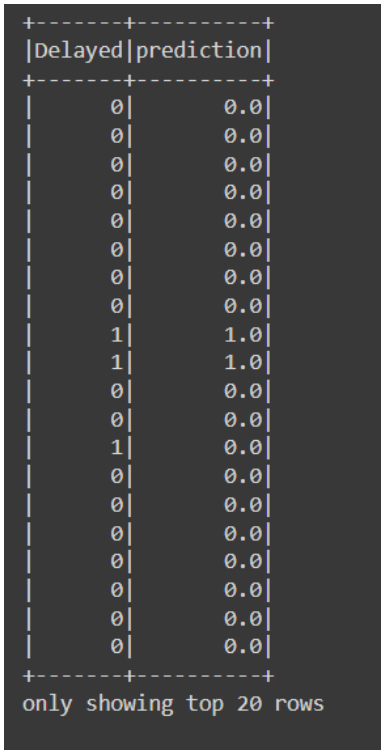
Logistic Regression

The training of Logistic Regression model was done with the help of PySpark's MLlib library, which is highly optimized for the machine learning algorithms. The model preprocessed flight data set, with specific flight attributes, such as time of departure, distance, as input variables and the target variable was the delay status, either delayed or not delayed. To improve the model's ability to generalize to unseen data cross validation was applied by dividing the dataset into training and test sets.

Random Forest Regression

Random Forest RegressioN was used with an aim of making predictions of the delay times in minutes. Delay time was again used as the target variable while the model was trained on the same preprocessed dataset as earlier. Random forest adopted in this research was flexible enough to handle non-linear relationships in the data and generate appropriate delay time forecasts.

Model Evaluation



Delayed	prediction
0	0.0
0	0.0
0	0.0
0	0.0
0	0.0
0	0.0
0	0.0
0	0.0
1	1.0
1	1.0
0	0.0
0	0.0
1	0.0
0	0.0
0	0.0
0	0.0
0	0.0
0	0.0
0	0.0
0	0.0

only showing top 20 rows

Finally when the model is trained, we used the MulticlassClassificationEvaluator for evaluating the accuracy of the model. The primary measure which was calculated was accuracy of the model, which is a percentage of correctly classified instances out of the total number of instances.

In this table:

- The **Delayed** column shows the actual flight delay status (1 for delayed, 0 for not delayed).
- The **Prediction** column shows the model's predicted status (1 for predicted delay, 0 for predicted non-delay).

The model achieved an overall accuracy of **84.9%**, indicating that the model correctly predicted flight delays in approximately 85% of the cases. While the model performed well, there were still instances of incorrect predictions, such as **false negatives** where delayed flights were predicted as non-delayed.

Here's a detailed explanation of the table:

- Each row shows a single observation from the test dataset.
- For example, in the first row, the actual value is 0 (no delay), and the model predicted 0.0 (correctly predicted no delay).
- Similarly, in the second and subsequent rows, the model consistently predicts the correct outcome for most of the observations, particularly for non-delayed flights.
- There are cases, like the 10th row, where the actual value is 1 (delay), and the prediction is also 1.0 (correctly predicting a delay).
- However, there are also incorrect predictions, as shown in row 13, where the actual value is 1 (delay), but the model predicted 0.0 (no delay).

This table demonstrates that the model performs quite well at predicting non-delayed flights, but there may still be some false negatives, where a flight delay was incorrectly predicted as no delay.

The Logistic Regression model provided a reasonably accurate prediction of airline delays, achieving an accuracy of 84.9%. This model can be further refined by tuning hyperparameters or exploring additional features that may impact flight delays. Future work could include experimenting with other classification models or ensemble methods to improve accuracy.

Random Forest Regression

```
Root Mean Squared Error (RMSE): 29.7149066082435
+---------+------------------+
|DEP_DELAY|        prediction|
+---------+------------------+
|     -5.0| 4.862127348797598|
|      1.0|2.7933636830864543|
|     -2.0|13.482310784023536|
|     -5.0| 4.678164223812877|
|     -4.0|13.918677240124799|
|     -2.0|0.6241488813807812|
|      3.0| 8.871183172274083|
|     -2.0|   6.8552168609644|
|    107.0| 20.81658021992305|
|    246.0|104.72646429989155|
|     -5.0| 6.206628516104127|
|     -4.0| 5.541540177771962|
|     16.0| 3.494042241401755|
|     -3.0| 3.411547882113028|
|      4.0|10.316829241232616|
|      6.0|13.918677240124799|
|     -3.0|13.679512296311179|
|     -3.0|16.013440732972143|
|     -8.0| 2.966682166401452|
|     -5.0| 8.629885762844454|
+---------+------------------+
only showing top 20 rows
```

For the Random Forest Regression model, the performance was evaluated using **Root Mean Squared Error (RMSE)**, which measures the average deviation between the predicted and actual delay times. The model achieved an RMSE of **29.71**, indicating that the model's predictions of delay times were, on average, 29.71 minutes away from the actual delay times.

The output table provides a comparison between the actual departure delays (DEP_DELAY) and the predicted delays (prediction) for a sample of 20 flights. Let's break this down:

- **DEP_DELAY**: This column shows the actual delay times of the flights in minutes. Negative values indicate that the flight departed earlier than scheduled, while positive values indicate delays.
- **prediction**: This column contains the predicted delay times generated by the Random Forest Regression model.

Here are some examples from the table:

1. **First Row**: The actual delay time is **-5.0** minutes (the flight departed 5 minutes early), and the predicted delay time is **4.86** minutes. The prediction is somewhat off, as the model predicted a slight delay, while the flight actually departed early.
2. **Tenth Row**: The actual delay is **246.0** minutes, and the model predicted a delay of **104.72** minutes. Although the model predicted a significant delay, it underestimated the actual delay.
3. **Ninth Row**: The actual delay is **107.0** minutes, and the predicted delay is **20.81** minutes. The model under-predicted the delay, showing that it might struggle with accurately predicting larger delays.

4. **Sixth Row**: The actual delay is **-2.0** minutes, and the predicted delay is **0.62** minutes. Here, the prediction is relatively close, as both the actual and predicted values suggest the flight was almost on time.

Insights and Model Performance

1. **Performance on Small Delays:** This can be said to be true because the model gives fairly good results especially on flights that had slight delays or those that were delayed shortly from the scheduled take off time. For instance, for the sixth row, it is dated as 6, whereas the actual delay was -2. Zero minutes, whereas the model was able to predict a minor delay of 0. It is quite close to the actual value; the authors estimated that it takes 62 minutes.
2. **Underestimation of Large Delays:** As with the majority of the flights with large delays, the model underestimates the actual time of the flight delay. This is evident from cases such as 10th row meaning, the actual delay being 246. 0 minutes, while the current model predicted only 104. 72 minutes. Such underestimations may suggest that the current model is not very effective in predicting extreme delay values.
3. **Handling Negative Delays:** This is because negative delays suggest that at least one vehicle set out early and thus there is more uncertainty. For example, in the first row, the actual delay was -5. 0 minutes (The flight actually departed a few minutes earlier than it was scheduled to do so) while the model predicted that the flight would be delayed by 4. 86 minutes. This implies that the model can still be finetuned a bit or might require new training data, specifically, with regard to people who check out early.

The **Random Forest Regression** model provided reasonable predictions for flight delay times but showed limitations in predicting extreme delays and early departures. The RMSE of **29.71** indicates a moderate level of accuracy, but the model could benefit from further tuning or the inclusion of additional features that could help predict extreme delays more accurately.

Challenges Faced During Model Training

1. **Imbalanced Dataset**: One of the challenges I faced during the training process was the imbalance in the dataset. Most flights in the dataset were not delayed, which caused the model to be biased toward predicting non-delay outcomes (0). This imbalance can lead to a higher accuracy but poor performance on minority classes (i.e., predicting delays).

2. **Feature Selection**: Selecting the right features for the classification model was challenging. Some features may not contribute to the model's predictive power, while others could have a

significant impact. Identifying and engineering the most relevant features was crucial in improving the model's performance.

3. **Model Tuning**: Tuning the hyperparameters of the Logistic Regression model to achieve optimal performance required several iterations. Finding the right combination of parameters (like the regularization strength and threshold) was challenging, as it significantly affected the model's accuracy.

4. **Data Preprocessing**: Handling missing values and ensuring that all features were properly encoded and normalized before feeding them into the model required careful preprocessing. Missing or incorrectly formatted data could have skewed the model's results.

5. **Overfitting**: During initial training phases, the model showed signs of overfitting, where it performed well on the training data but not as well on the test data. To overcome this, I applied techniques like cross-validation and regularization to prevent overfitting and ensure the model generalizes well to new data.

6. **Class Imbalance Evaluation Metrics**: Relying solely on accuracy as an evaluation metric proved insufficient, especially in an imbalanced dataset. While the model achieved 84.9% accuracy, the precision and recall for delayed flights needed further evaluation to ensure the model wasn't biased toward the majority class.

DATA ANALYSIS AND DATA VISUALIZATION

Oecumenical big data analysis workflow.

Below workflow diagram emphasizes the end-to-end big data analysis workflow from data acquisition to usable key insights and highlighted the analyzing the large datasets. Mainly this applied Microsoft Azure services to streamline each stage of this process. Eventually preprocessed data integrated with Power BI to absorb the extract valuable insights and drive business decisions Each step have been described summarized manner.

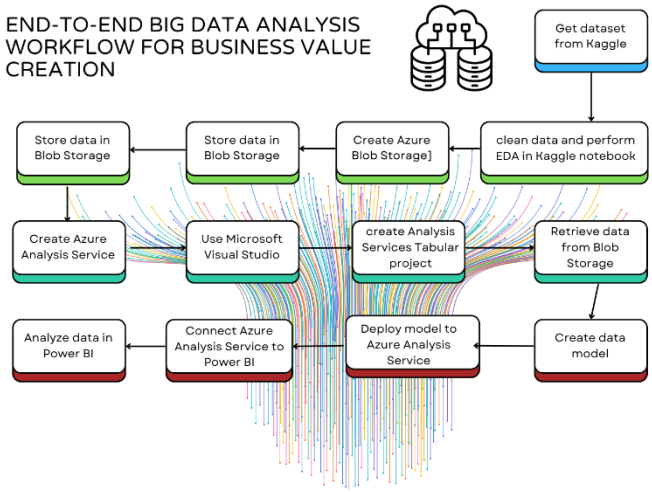


Fig no. 01- Workflow diagram for overall big data analysis

1. Selecting the overall consistent dataset from Kaggle.
2. Preprocess the data and investigating an explanatory variable analysis in order to identify the relationship between categorical variables and continuous variables.
3. Create a MS Azure storage account and configure a container to store the blobs.

   Firstly, entering the storage account tab, wants to fill out project details and instance details. Subscription and resource group was respectively “Azure for student” and “(New) BI. Created “biairline” as the storage account name apart from that filled the region, primary service, performance and redundancy areas. The proof in the below image,

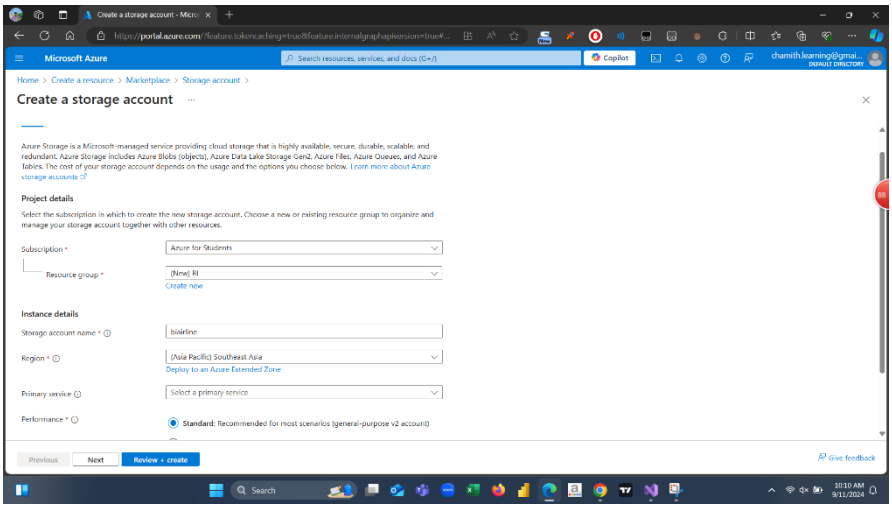


Fig no. 02. Creating storage account

4. Store data in blob storage.

In this stage have been selected the existing container which is “airlinecleaned” and upload the files in it and there will be a unique URL appeared.

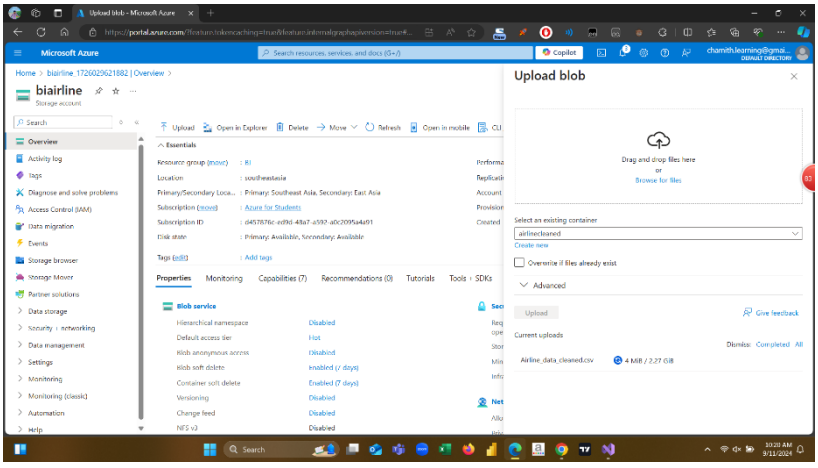


Fig no 03. Store data in blob

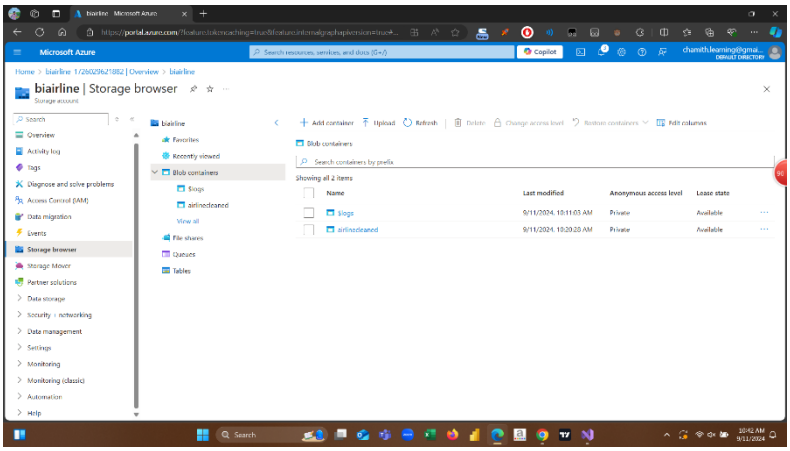


Fig no 04. Blob container

5. Creating an Azure Analysis Services

This platform facilitates models like tabular and multidimensional and it is easy to combine with powerful visualization tools like Power BI, Tubule, Excell and SQL. After creating an Azure Analysis Services need to configure, there are couple of details options, such as subscription, resource group, region, server name, pricing tier. The need to connect to Azure Analysis Services with SQL Server Management Studio (SSMS) or Visual Studio.

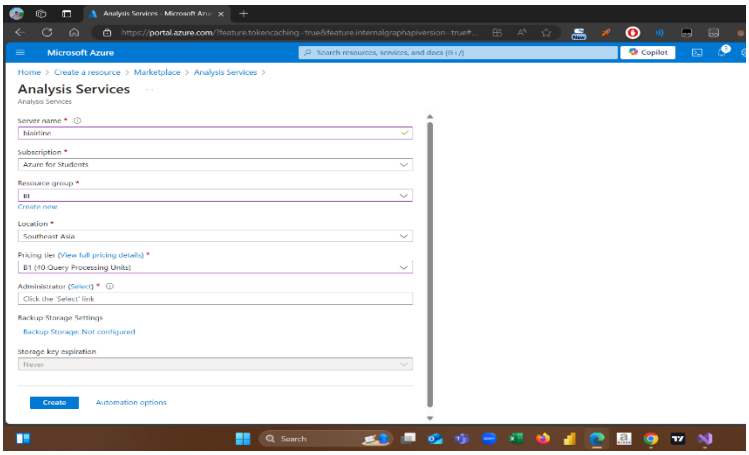


Fig no. 05 Creating the Azure Analysis Service

6. Using MS Visual Studio.

   To integrate with Azure Analysis Services models in the environment of Visual Studio, there are some extensions, in Analysis Services Projects. After installing the extensions to the Microsoft Analysis Services Projects. Subsequently, may need to restart Visual Studio.

7. Creating Analysis Services Tabular projects.

   After establishing an Analysis Tabular Projects Services Setting up a data model using the Tabular data model in Microsoft Visual Studio is necessary for creating interactive reports and analytical solutions. Firstly, selecting the Analysis Services Tabular Project there are settings to configure such as project name, location, solution name, framework. After filling those details screen is visible as follow,

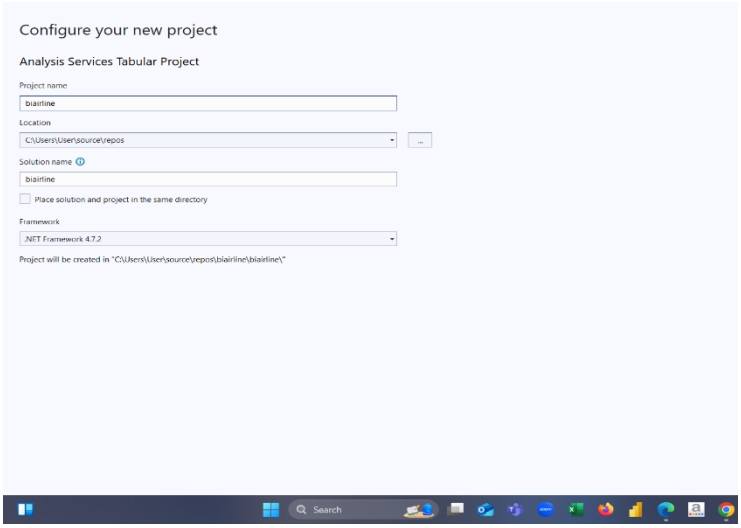


Fig no. 06 Create the new tabular project

Next, should need to configure the data source so after creating the new tabular project there will be a default Tabular model. So, in that need to be choose integrated workspace which eliminates the need to provide an explicit AS server instance. In addition, choosing the type of data source, wants to connect and entering the connection details is a must. Afterward, configuring the connection select the tables that wants to build the model.

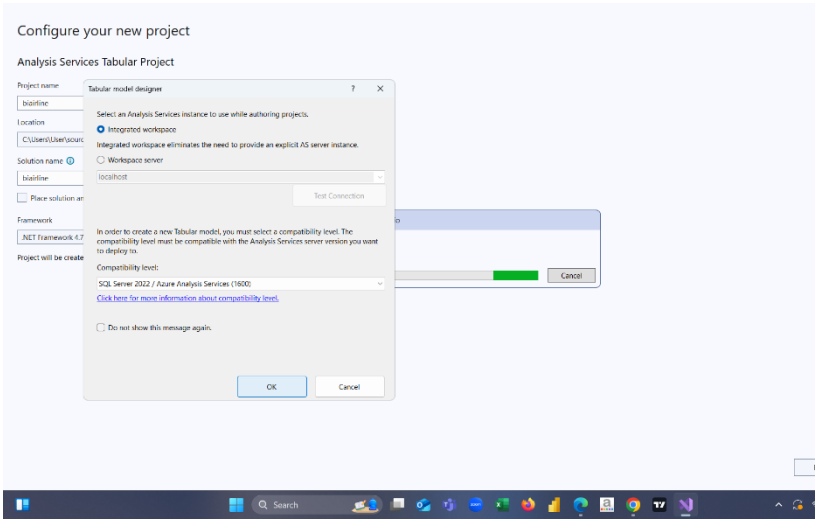


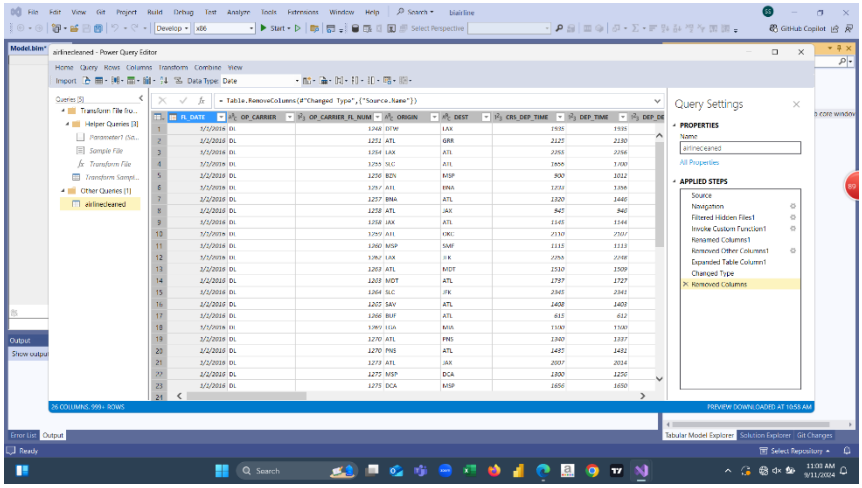
Fig no. 07 Configure data sources

8. Retrieve data from blob storage.

Retrieving data from blob storage is a crucial step firstly navigate to Azure Blob storage and select the container. After the container that will hold the blob and be able to retrieve it. After the screen is visible the related properties.

9. Creating the model

In this stage in the Tabular model explorer there were some applied steps, like navigation, filtered some files, invoke custom functions, rename and remove other unnecessary columns, expand the table column, change the type like modeling steps are done.



Next, evolve the configure the processing progress, so processing gets updated data from the original data sources.

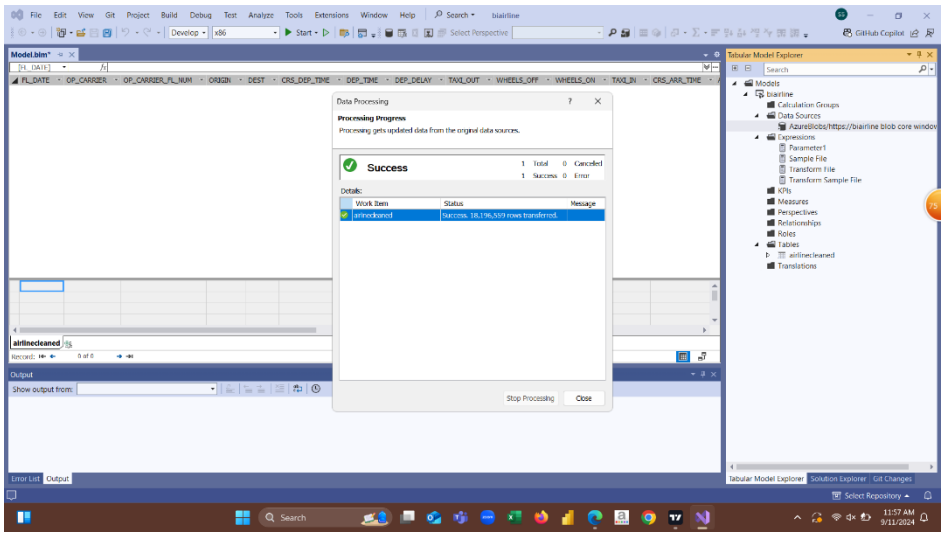
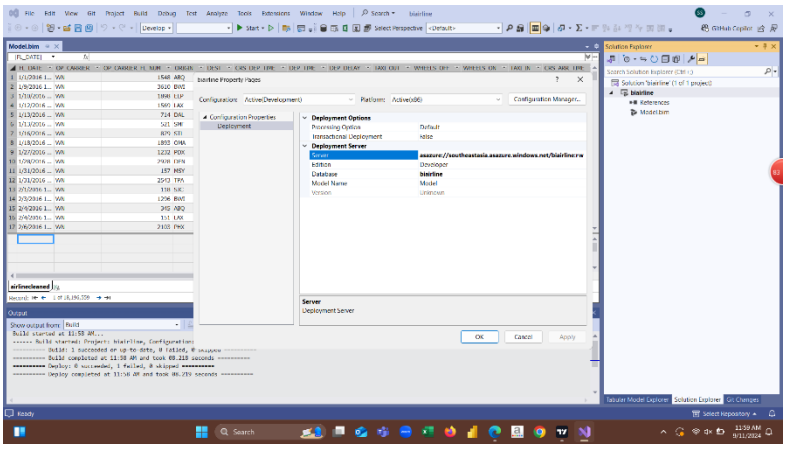
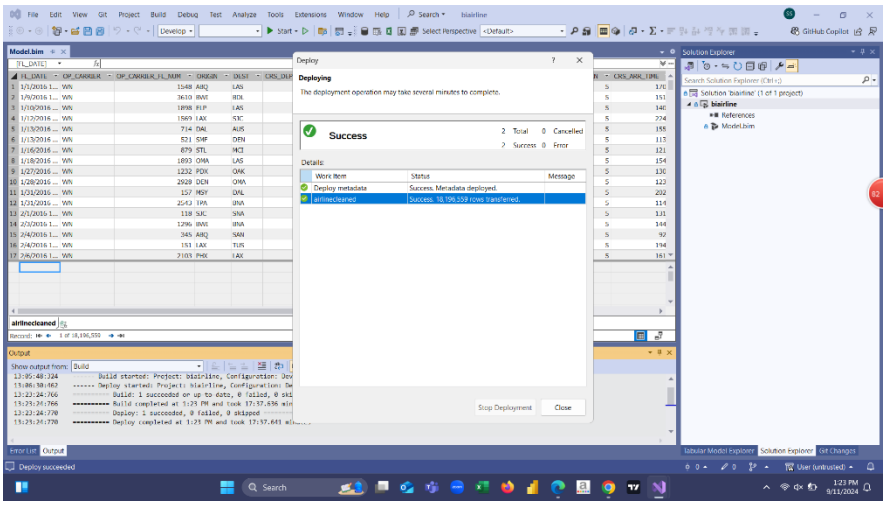


Fig no. 09 Configure the data processing progress

10. Deploy model to Azure Analysis Service.

    Deploying a model to Azure Analysis Services which has several steps to do, and this has been done from MS Visual Studio using Analysis Services Projects extension. After preparing the model in Visual Studio need to fully configure and move to configure deployment settings. For that connecting the created project and select properties in solution explorer and there are deployment settings such as server name, database name that should be filled. Afterward moving to the deploy the model, there will be a deployment process and output window in Visual Studio will show the progress as discussed earlier. If there are any issues that will also appear in the window. Eventually verify the deployment.





11. Connect Azure Analysis Service to Power BI

    Basically, Azure Analysis Service connected to the Power BI desktop for the after opening it, moving to get data option and by searching the Azure Analysis Services and connect it. Then entering the server name is necessary, after connecting Power BI will show a list of available database.

12. Analyze data in Power BI.

    This step is broadly discussed in dashboard analysis.

The key insights and objectives of the dashboard.

Basically, in these dashboards focuses on main 3 objectives that we identified, Analysis those and what are the key insights, patterns, trends, relationships, predictions and inferences which will be a very crucial aspects for the Airlines, as well the other influence companies and fields. Such as logistic and cargo sector, catering and food services, ride hailing and ground, hospitality industry and more. The objectives are,

- Understand the cause and patterns of the flight delays

Mainly the dashboards are able to identify the root causes by analyzing delay reasons, time of delay, weather conditions which focus to elaborates clear overview of how it influences by other factors.

- Assesses airline performance

  The dashboards are useful for comparing the punctuality records of various airlines, spotting anomalies, and pinpointing areas where their operations need to be improved.

- Evaluate operational efficiency

  The dashboards are able to assist in identifying inefficiencies and places where changes may be made to minimize delays by analyzing data on airport operations, taxiing times, and other pertinent indicators.

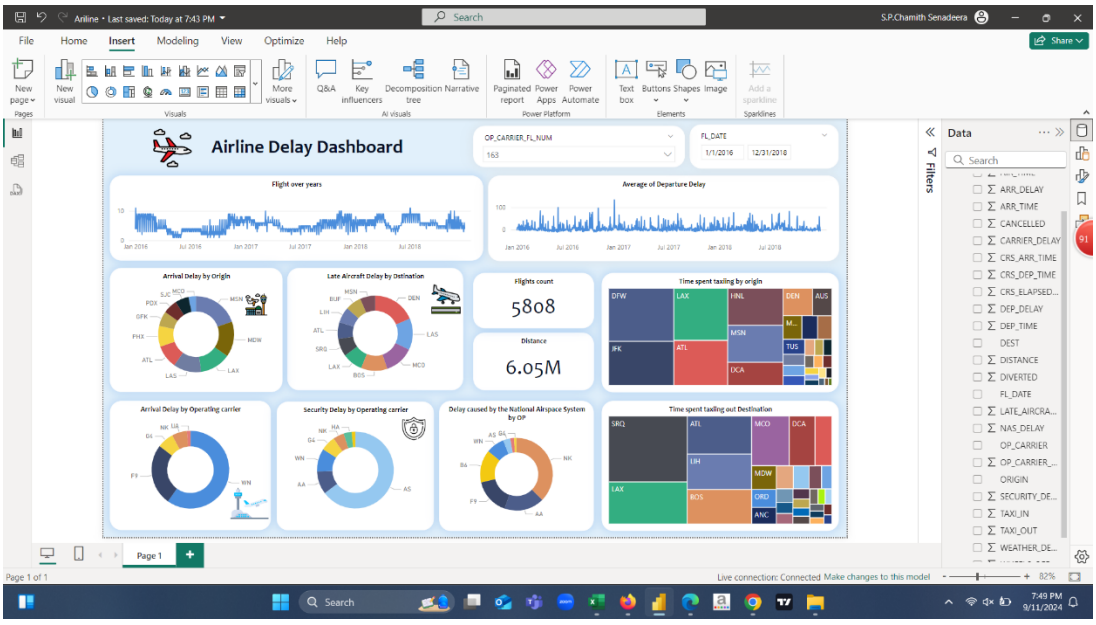


Fig no. 12 Dashboard 01

➢ **Key insights**

This dashboard is mainly aims from 2016 – 2019 timeline.

- Flight over years – There is no considerable high or low fluctuations, seasonal fluctuation can be seen.
- Average of departure delay – There are some peaks with certain time frame of increased delay.

- Flight statistics – 5808 of flight count and 6.05 million miles of total distance are covered those parameters emphasize the scale.
- Arrival delay by origin – This pie chart focuses on the delays by the city of origin, there are notable cities like Madison (MSN), Midway (MDW), Los Angeles (LAX), Las Vegas (LAS) proportionally high.
- Late aircraft delay by distance – This pir chart focuses brilliantly, the delays by destination. Respectively high Denver (DEN), Las Vegas (LAS), Orlando (MCO), Boston (BOS).
- Arrival delay by operating career – This pie chart visualizes the most airline delays on operating career. There is a significant amount of portion is captured by Southwest Airlines (WN) in addition Frontier Airlines (F9).
- Security delay by operating career – On the one important parameter security delay by operating career significant portion is captured by Alaska Airlines (AS). Apart from that American Airlines (AA), Southwest Airlines (WN) proportionally same portion is captured.
- National airspace system by OP – Based on this parameter Spirt Airlines (NK), American Airlines (AA), Frontier Airlines (F9), JetBlue Airlines (B6) are reported more than others.
- Time spent taxing by origin – Respectively DFW, JFK, ATL, LAX, HNL, HNL are represented proportionally.
- Time spent taxiing out destination - ATL, MCO, and DCA are important destinations with longer taxiing times, while DFW, LAX, and JFK are popular for taxiing from the origin side.

Airport operations and airline management as well influenced other parties can gain useful information by using the pie charts and tree maps to pinpoint severe delays and taxiing times.

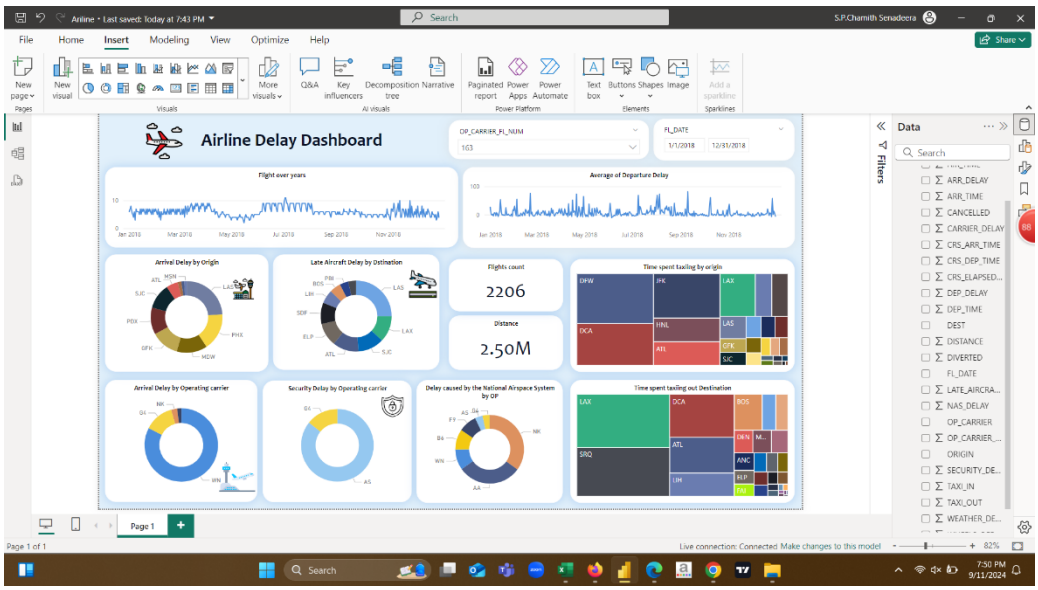


Fig no 13. Dashboard 02

- **Key insights**

  This dashboard is mainly aims from 2018 – 2019 timeline.

  - Flight over years – There is a sight fluctuation with the fight amount, moreover some peaks and troughs.
  - Average of departure delay - There are prominent peaks, which point to times when departure delays were noticeably longer than average, they could be caused by operational or seasonal causes.
  - Flight statistics – 2206 of fight count and 2.50 million miles have been travelled in total demonstrating the operating reach in terms of miles.
  - Arrival delay by origin – San Josa (SJC), Los Angeles (LAX), Atlanta (ATL) are in the chart there is a high proportion is represented. This might point to problems with those origins' operations.
  - Late aircraft delay by distance – Palm Beach International (PBI), Los Angeles (LAX), Atlanta (ATL) have higher late aircraft delays and that might be inbound flights or airport congestion.
  - Arrival delay by operating carrier - In terms of delays, carriers like WN and NK stand out, indicating which airlines may see more operational delays.
  - Security delay by operating carrier - Airlines are a good way to illustrate security related delays, carriers like AA and AS seem to have higher delays as a result of security checks or associated procedures.
  - National airspace system by OP - It seems that delays associated with the National Airspace System which includes traffic control and airspace regulations affect airlines like F9, AS, and WN more than others.
  - Time spent taxing by origin - At DFW, JFK, and LAX, significant taxiing durations are seen, suggesting potential congestion or longer runway distances at these airports.
  - Time spent taxiing out destination - The longest taxiing times are at airports like DCA, BOS, and ATL, suggesting possible problems with airport operations at these sites.

EFFECT OF DATA ANALYSIS ON BUSINESS VALUE

Selected company

For this assignment, we have chosen a Logistics and Cargo Company as the focal point of our data analysis. This company operates within the logistics industry, providing a wide range of services including cargo transportation, warehousing, inventory management, and supply chain solutions. The company's primary goal is to streamline the movement of goods across various locations while ensuring timely delivery, cost efficiency, and customer satisfaction.

The company caters to both domestic and international markets, handling shipments across different modes of transportation such as air, sea, and land. Its clientele includes businesses from diverse industries like retail, manufacturing, and e-commerce, all of which rely heavily on the efficient movement of goods to maintain their operations. As a result, the company deals with high volumes of data, including shipment details, route optimization, delivery timelines, and resource management, all of which are vital for ensuring smooth logistical operations.

Given the importance of data in logistics, our analysis focuses on optimizing key areas such as supply chain efficiency, demand forecasting, and cost reduction. The insights generated through our analysis will provide the company with strategies to improve resource allocation, reduce delays, minimize operational costs, and enhance overall service delivery.

By leveraging advanced data analysis and cloud-based technologies, the Logistics and Cargo Company aims to achieve a competitive edge in the industry, ensuring that it remains adaptable to changing market conditions while continuing to meet customer demands efficiently.

How important the above-mentioned key insights to the Logistic and Cargo company

For logistics and cargo corporations, the insights from the two Power BI dashboards that were created to analyze aircraft delays are highly helpful, especially when it comes to routing optimization, improving delivery times, and reducing risks. The importance of each insight in assisting cargo firms in achieving more dependable and efficient operations has been covered in the sections below.

- Flight trends.

  Logistics planners and cargo companies become aware of potential interferences throughout the course of the year, during festive time or months with unfavorable weather since they are capable of observing trends of airline delays and fluctuations in Monthly Traffic. They can

arrange other routes, obtain more capacity, or change the frequency of shipment so that they can avoid being affected by such fluctuations.

- The average of departure delay.

  Flight delays indicate that there are constraints that may possibly slow down delivery of CSD cargo aircrafts affecting the deliveries of consignments at the end of the flow. That is why carriers and operators, with the help of such information, can either add extra time into schedules or switch to other operators or times of the day or day of the week, staying away from peaks that consist of substantial increases in departure delay frequencies caused by operations or seasons.

- Flight statistics

  This is evident by the numerous flights and total distance that has been covered by the freight businesses in order to ensure that they have met their intended targets. Such figures show the need to allocate for incidental costs to ensure that General Motors continues to adhere to precise flight schedules especially for flights that take long to get to the next destination of General Motors' supply chain.

- For the dependent variables; arrival delays by origin

  Cargo operators can avoid some of these flights coming from particular starting points including Los Angeles (LAX), Madison (MSN), and Las Vegas (LAS) depending with the knowledge that some arrival airport often experience delay. This is especially useful for products that are time-sensitive, that is products, that should reach their destination as early as possible. This helps businesses to ensure timely delivery schedules and at the same time fully consider other routes that possibly may not be frequently delayed.

- Late aircraft delays by destination.

  Operations related to cargo transfer can be seriously hampered by late flight delays, especially at busy airports like Denver (DEN), Las Vegas (LAS), and Orlando (MCO). Knowing these locations enables cargo companies to improve reliability by rerouting shipments through less crowded airports or allocating more time for transfers in their logistics timetables.

- Arrival delays by operating carrier.

  Due to their reputation for frequent delays, carriers such as Southwest Airlines (WN) and Frontier Airlines (F9) should be avoided by cargo companies when making reservations.

Logistics companies can more effectively guarantee on-time delivery by avoiding airlines that have a history of delays or by establishing backup plans for shipments with these carriers.

- Security delays by operating cartier
  Security checks on cargo are essential, but security process delays can be harmful. For instance, the percentage of security-related delays on Alaska Airlines (AS) and American Airlines (AA) is higher. This suggests that a cargo company may favor carriers with lower security-related delays, particularly for important or valuable cargo. This reduces the possibility that cargo will become stranded during drawn-out security inspections.

- National airspace system delays by operating carrier.
  Although they can be challenging to manage, delays in the National Airspace System caused by air traffic control or airspace restrictions are a big worry for carriers such as Spirit Airlines (NK), Frontier (F9), and JetBlue (B6). Avoiding routes or carriers that are disproportionately impacted by NAS delays is essential for cargo logistics because these delays frequently worsen and impact delivery schedules.

- Time spent taxiing by origin and destination.
  Airports with lengthy cab waits for both arrivals and departures include DFW, JFK, ATL, and LAX. The total time it takes to transfer freight can be greatly impacted by longer taxiing times, particularly for deliveries that need to be made on time. Cargo businesses can choose airports with shorter turnaround times by knowing which ones have longer taxiing times. This helps minimize idle time on the tarmac.

As overall logistic and cargo companies who might be these insights as indicators,

- Improved route planning
- Carrier selection optimization
- Buffering and contingency planning
- Seasonal and peak time adjustments
- Operational efficiency
- Cost saving

CHALLENGES FACED

The completion of this assignment presented several significant challenges, each contributing to the complexity and demanding nature of the task. Here's a comprehensive breakdown of the hurdles faced and how they impacted the assignment:

1. Selecting an Appropriate Dataset:

The first task was therefore to identify a data set that was meaningful and informative but also fit the task description at hand. It was important as all the subsequent analysis and learning was built upon this step.

2. Handling Large Data Volumes:

According to the chosen dataset it was vast, which caused delays in the time taken to download the data and loading of the data. Large amount of data also increased the difficulty in data cleaning and preprocessing steps and increased time and computational power needed to resolve them.

3. Big Data Management:

Challenges: The most difficult issue here was to manage big data within the scope of this assignment. In order to manage and analyze such large volumes of data there was a need to optimize the handling of data which was not easy to do.

4. Learning and Utilizing Azure:

It involved the applicability of the Azure platform which is a complicated cloud computing service. This required a learning slope all by itself in the sense that a basic understanding of the tools and processes that Azure provides was critical in how it could be used to optimize the project.

5. Software and Infrastructure Limitations:

There was use of heavy software throughout the assignment especially heavy tools such as Visual Studio. The capacity in most of these structures was sometimes inadequate to support these computationally demanding applications which thus slowed down work processes.

6. Budget Constraints with Azure:

There is also the problem of finance bottlenecks in the usage of Azure as well as some other issues. Initially, the project was possible due to a free trial, in which each user had a $100 limit and could use it only in two days. This limited working time made the project to be compacted within a given period of time and once the trial runs expired, all developments made within the said platforms are erased, this is the nature of resorting to limited trial services.

7. Overall Complexity Due to Big Data:Overall Complexity Due to Big Data:

The common theme of the challenges posed into account was the issue of data size, referring to big data. Most of the aspects realized during the completion of the assignment were shaped by requirements of working with points, which are numerous in big data.

These challenges were not limited to testing the aptitudes and robustness of the data analytic team but also taught the principles of planning flexibility in the analytic projects. Each challenge was informative to the process demonstrating that big data work in practice has many intertwined aspects.

CONCLUSION

In conclusion, this assignment showed how the principles of big data analytics can be applied and yield business intelligence for a Logistics and Cargo company. Thanks to the use of additional tools, such as PySpark, Azure, and Power BI, we were able to process big data and avoid various problems associated with data preparation and missing values as well as with the training of models. The research applied the Logistic Regression and Random Forest Regression to formulate models that presented essential features that affect airline delays and cancellations and provided recommendations to enhance the performance of the airline.

These observations are particularly pertinent to the logistics industry as they reveal patterns and problematic aspects of airline management that may have an impact on supply chains. In this way, decision-making based on business analytics can have a positive impact on lean business operations and lead to better delivery times, improved customer satisfaction, cost reduction for businesses, which will result in better competitiveness overall.

While there were some challenges regarding the integration of big data for analysis work due to big data characteristics, software issues, and some limitations of infrastructure, we were able to showcase how big data technology can revolutionize business operations and transform decision making in the organization.