Airline Delays Prediction

Mustafa Aydoğan
Computer Engireening
TOBB ETÜ
Ankara, Turkey
m.aydogan@etu.edu.tr

***Abstract*—This project is a machine learning project that trains a model that calculates the probability of flight delays and take advantage of this model.**

***Index Terms*—airline, delay, prediction**

I. Introduction

Airplanes, which are accepted as the fastest means of transportation today, allow us to reach the places we want to go in less time. However, due to weather conditions and some technical problems, the planes may be delayed. These delays can last 15 minutes or more. Passengers boarding the plane may not care about delays of less than 15 minutes, but they will be annoyed by delays of more than 15 minutes. Passengers always wonder if their flights will be delayed. We want to train a model that calculates the probability of flight delays and take advantage of this model. We will use a dataset obtained from Kaggle while training this model. First of all, we will read the data from the file and organize it according to our needs. We will identify the required attributes and make the data as simple as possible. We will divide the data we have into train, validation and test. We will then train our model. We will validate our model using validation data. Then we will evaluate it with test data. We will compare the results from our model with the test data and try to measure the performance of our model. Our aim is to train a model capable of binary classification that calculates whether there will be a delay of at least 15 minutes in flights. Although, we will evaluate the results comparing the model results and test data.

II. Literature Review

A. Flight Delays and Cancellations

Airline delays and cancellations can be a major inconvenience for passengers, and can also have a significant impact on the airline industry as a whole. There are a number of factors that can contribute to airline delays and cancellations, including:

- Weather: Weather is the most common cause of airline delays and cancellations. Bad weather conditions, such as thunderstorms, snow, and fog, can make it difficult or impossible for planes to take off and land.
- Technical issues: Technical issues can also cause airline delays and cancellations. For example, if a plane has a mechanical problem, it may need to be repaired or replaced before it can take off.
- Staffing issues: Staffing issues can also contribute to airline delays and cancellations. If there are not enough pilots, flight attendants, or other staff members available, flights may need to be delayed or canceled.
- Air traffic control: Air traffic control can also cause airline delays and cancellations. If there is too much air traffic in a particular area, flights may need to be delayed or canceled.
- Security: Security checks can also cause airline delays and cancellations. If there is a security threat, flights may need to be delayed or canceled.
- Strikes: Strikes by airline employees can also cause airline delays and cancellations. For example, if pilots or flight attendants go on strike, flights may need to be delayed or canceled.
- Political unrest: Political unrest in a particular region can also cause airline delays and cancellations. For example, if there is a war or civil war in a particular country, flights to and from that country may be delayed or canceled.
- Natural disasters: Natural disasters, such as earthquakes, hurricanes, and floods, can also cause airline delays and cancellations. For example, if there is a major earthquake in a particular region, flights to and from that region may be delayed or canceled.

Other factors There are a number of other factors that can contribute to airline delays and cancellations, such as:

- Airport congestion
- Bad weather at the destination airport
- Medical emergencies
- Mechanical problems with baggage handling equipment

III. Dataset, Data properties, Attributes

1) **Dataset Source:** https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations
2) **The dataset contains the following columns:**
    - *MONTH = Month*
    - *DAY_OF_WEEK = The Day of Week*
    - *DEP_DEL15 = TARGET Binary of a departure delay over 15 minutes (1 is yes)*
    - *DEP_TIME_BLK = Departure time block*
    - *DISTANCE_GROUP = Distance group to be flown by departing aircraft*
    - *SEGMENT_NUMBER = The segment that this tail number is on for the day*

- *CONCURRENT_FLIGHTS = Concurrent flights leaving from the airport in the same departure block*
- *NUMBER_OF_SEATS = Number of seats on the aircraft*
- *CARRIER_NAME = Carrier*
- *AIRPORT_FLIGHTS_MONTH = Avg Airport Flights per Month*
- *AIRLINE_FLIGHTS_MONTH = Avg Airline Flights per Month*
- *AIRLINE_AIRPORT_FLIGHTS_MONTH = Avg Flights per month for Airline AND Airport*
- *AVG_MONTHLY_PASS_AIRPORT = Avg Passengers for the departing airport for the month*
- *AVG_MONTHLY_PASS_AIRLINE = Avg Passengers for airline for month*
- *FLT_ATTENDANTS_PER_PASS = Flight attendants per passenger for airline*
- *GROUND_SERV_PER_PASS = Ground service employees (service desk) per passenger for airline*
- *PLANE_AGE = Age of departing aircraft*
- *DEPARTING_AIRPORT = Departing Airport*
- *LATITUDE = Latitude of departing airport*
- *LONGITUDE = Longitude of departing airport*
- *PREVIOUS_AIRPORT = Previous airport that aircraft departed from*
- *PRCP = Inches of precipitation for day*
- *SNOW = Inches of snowfall for day*
- *SNWD = Inches of snow on ground for day*
- *TMAX = Max temperature for day*
- *AWND = Max wind speed for day*

3) **Preprocessing:** Our model was overfit because there were not equal numbers of true and false values in the data set. Therefore, a preprocessing was done to make the data count balanced.
4) **Using Models:**
This part is about the different models we used for our project. We split our dataset into two parts: one for training (67% of the data) and the other for testing (33%). This helps our models learn patterns in most of the data and then see how well they perform on new, unseen data.
We mostly focus on two choices for classifying flights: if they're ”delayed” or ”not delayed.” This helps us tackle the main issue of predicting flight delays.
Here are the models we picked:

   a) **Logistic Regression:** Think of this like a tool to predict a simple yes or no. It learns from data and figures out the probability of a flight being delayed or not. We'll use this to fine-tune its learning process and make accurate predictions.
   b) **Decision Tree:** Imagine a flowchart that helps decide things. That's what a Decision Tree is. It looks at different factors and follows a path to classify flights as delayed or not based on the history of other flights.
   c) **GaussianNB (Naive Bayes):** This one is like a detective. It calculates the probability of a flight being delayed using math and statistics. It's called ”Naive” because it assumes the different factors are not connected to each other, which simplifies things.
   d) **MLPClassifier (Multi-Layer Perceptron):** This is like a bunch of detectives working together. It's a bit more complicated but good at figuring out patterns even when they're not obvious. It can handle complex relationships between different factors.
   e) **RandomForestClassifier:** This is like a team of decision-making trees, where each tree has its say, and then they all vote to make a final prediction. This teamwork ensures accuracy and is great for complex tasks, like predicting flight delays.

We picked these models because they're known to work well for sorting things into categories, which is what we're doing with flights. They each bring something different to the table, making our predictions more accurate. In the next parts, we'll explain how we used these models, trained them, and show the results to see which one works best for predicting flight delays.

IV. Progress

A. What have we done? What will we do?

First, we downloaded the dataset from Kaggle and checked for missing data. Then, we converted the nominal data into numerical data for effective analysis. We employed visualization techniques to gain insights into the relationships between attributes. Once we had a clear picture, we normalized the data to ensure uniform scaling.

After normalizing the dataset, we proceeded to split it into training and testing sets. To identify the most suitable model for our task, we trained and evaluated several classifiers including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gaussian Naive Bayes, and Multi-Layer Perceptron Classifier. We aimed to determine the model that performs best for our dataset and problem.

To further enhance our approach, we leveraged Principal Component Analysis (PCA) for dimensionality reduction. By applying PCA, we reduced the complexity of our feature space while retaining essential information. This enabled us to manage computational resources more efficiently and reduce the potential for overfitting.

In the next stage, we carried out feature selection to identify the most influential attributes. After identifying these key features, we streamlined the dataset to focus on them. With this refined dataset, we retrained the selected model using PCA-transformed features. This iterative process allowed us to refine our model's accuracy by harnessing the power of PCA and optimizing feature selection.

Ultimately, our goal is to identify the optimal combination of preprocessing techniques, model selection, and feature engineering to create a high-performing predictive model tailored to our specific problem.

V. Results and Model Discussion

Our journey through the world of airline delay prediction has culminated in the evaluation of various classification models. Each model brings its own unique approach to the table, and the results they yield shed light on their strengths and limitations in tackling the challenge of predicting flight delays.

1) **Logistic Regression:** Logistic Regression, our initial choice, demonstrates a respectable accuracy of approximately 0.84 on the test dataset. It efficiently makes binary predictions and excels in scenarios where the relationship between features and outcome is linear. However, in the context of flight delays, where factors can interact in complex ways, its performance might be limited.
2) **Decision Tree Classifier:** The Decision Tree Classifier impresses with a test accuracy of around 0.85. Its ability to create intuitive decision paths makes it suitable for interpreting the model's behavior. However, decision trees are prone to overfitting, which might lead to reduced generalization on unseen data.
3) **Gaussian Naive Bayes:** Gaussian Naive Bayes offers a test accuracy of about 0.83. It's particularly useful when dealing with large feature spaces and assumes that features are conditionally independent given the class label. Although this simplifying assumption might not always hold true in real-world scenarios, the model still delivers a commendable performance.
4) **Random Forest Classifier:** The Random Forest Classifier, a collection of decision trees, achieves a test accuracy of roughly 0.85. Its ensemble nature enhances generalization and robustness, making it less susceptible to overfitting. Random Forests are capable of capturing complex relationships within data, which aligns well with the intricate nature of flight delays.
5) **MLPClassifier (Multi-Layer Perceptron):** The MLPClassifier stands out as a frontrunner, boasting a test accuracy of around 0.86. Its deep neural network architecture allows it to learn intricate patterns and relationships within data, making it well-suited for capturing the complex factors contributing to flight delays.

In our endeavor to predict flight delays, we have traversed the realms of data preparation, model selection, and performance evaluation. The culmination of our efforts rests in the results we have obtained and the insights we have gained from this process.

We began by tackling the inherent imbalance in our dataset, as a skewed distribution could lead to biased predictions. Through careful preprocessing, we balanced the dataset, ensuring that both delayed and non-delayed flights were represented equally. This preprocessing step laid the foundation for accurate model training and evaluation.

With our dataset ready, we embarked on the exploration of various classification models. We introduced our models, each with its unique approach to understanding and predicting flight delays. Logistic Regression, akin to a binary decision maker, Decision Trees acting as logic flowcharts, Gaussian Naive Bayes employing probabilistic insights, Multi-Layer Perceptron as a network of pattern detectors, and the collaborative power of RandomForest were all brought into play. By testing these models against our dataset, we aimed to identify the champion in predicting flight delays.

The results of our model evaluation revealed that the MLPClassifier emerged as the frontrunner, showcasing an accuracy of approximately 0.86. This model's adeptness in capturing intricate relationships within the data proved invaluable for predicting flight delays accurately. The journey, however, did not stop here; we were determined to refine our approach further.

Incorporating the principles of Principal Component Analysis (PCA), we delved into dimensionality reduction, a technique that helps streamline data while preserving its essential characteristics. By employing PCA, we sought to enhance computational efficiency without compromising prediction accuracy. The amalgamation of the **MLPClassifier** with PCA-transformed features led to a model that embraced the twin strengths of advanced classification and dimensionality reduction.

VI. Conclusion

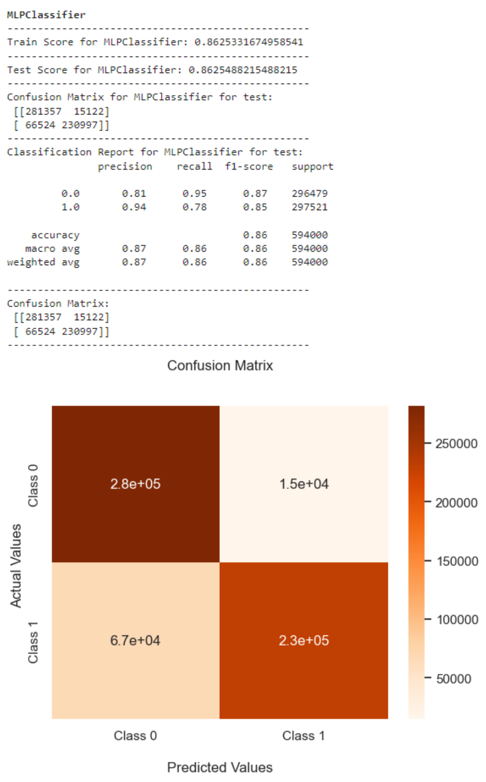
In the tapestry of airline delay prediction, our journey has been one of exploration, innovation, and determination. Fueled by the aspiration to empower passengers and the aviation industry with insights, we ventured into the world of machine learning armed with data and curiosity.

We've emerged from this journey with a model that predicts flight delays with a remarkable accuracy of approximately 0.86, and the power of dimensionality reduction through PCA has been harnessed to enhance efficiency. However, the impact of our voyage extends beyond these numbers. We've unearthed the complexity of flight delays, unveiled patterns hidden within data, and leveraged technology to navigate this intricate landscape.

As we conclude this chapter, we recognize that our journey continues. The skies above are ever-changing, and our commitment to enhancing the travel experience is unwavering. With each prediction, we inch closer to a seamless, efficient, and passenger-centric aviation industry. Our story is one of data-driven innovation, and its next chapter holds the promise of progress, precision, and positive change.

VII. Data Visualition

In this section, results of each models, confusion matrix and correlation matrix are visually available. In this way, we can comment and discuss in a better way.



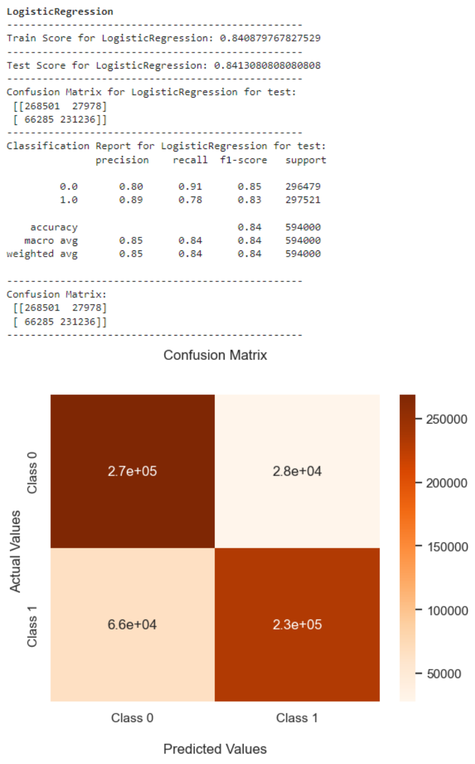
```
MLPClassifier
--------------------------------------------------
Train Score for MLPClassifier: 0.8625331674958541
--------------------------------------------------
Test Score for MLPClassifier: 0.8625488215488215
--------------------------------------------------
Confusion Matrix for MLPClassifier for test:
 [[281357  15122]
 [ 66524 230997]]
--------------------------------------------------
Classification Report for MLPClassifier for test:
               precision    recall  f1-score   support

         0.0       0.81      0.95      0.87    296479
         1.0       0.94      0.78      0.85    297521

    accuracy                           0.86    594000
   macro avg       0.87      0.86      0.86    594000
weighted avg       0.87      0.86      0.86    594000

--------------------------------------------------
Confusion Matrix:
 [[281357  15122]
 [ 66524 230997]]
--------------------------------------------------
```

Fig. 1. MLPClassifier



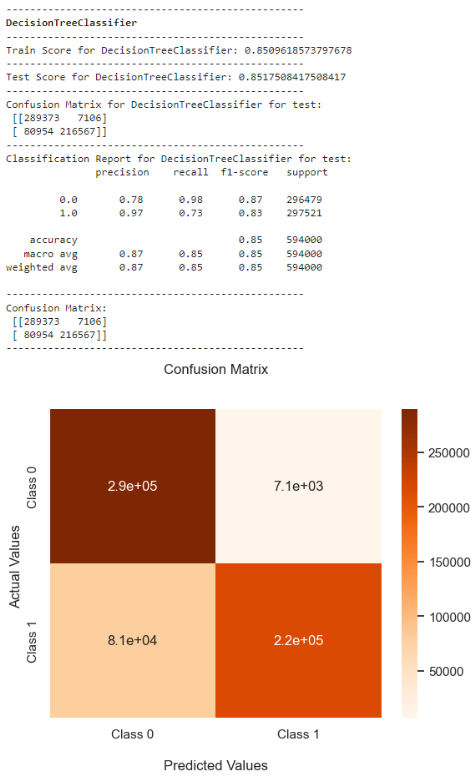
```
LogisticRegression
--------------------------------------------------
Train Score for LogisticRegression: 0.840879767827529
--------------------------------------------------
Test Score for LogisticRegression: 0.8413080808080808
--------------------------------------------------
Confusion Matrix for LogisticRegression for test:
 [[268501  27978]
 [ 66285 231236]]
--------------------------------------------------
Classification Report for LogisticRegression for test:
               precision    recall  f1-score   support

         0.0       0.80      0.91      0.85    296479
         1.0       0.89      0.78      0.83    297521

    accuracy                           0.84    594000
   macro avg       0.85      0.84      0.84    594000
weighted avg       0.85      0.84      0.84    594000

--------------------------------------------------
Confusion Matrix:
 [[268501  27978]
 [ 66285 231236]]
--------------------------------------------------
```

Fig. 2. Logistic Regression



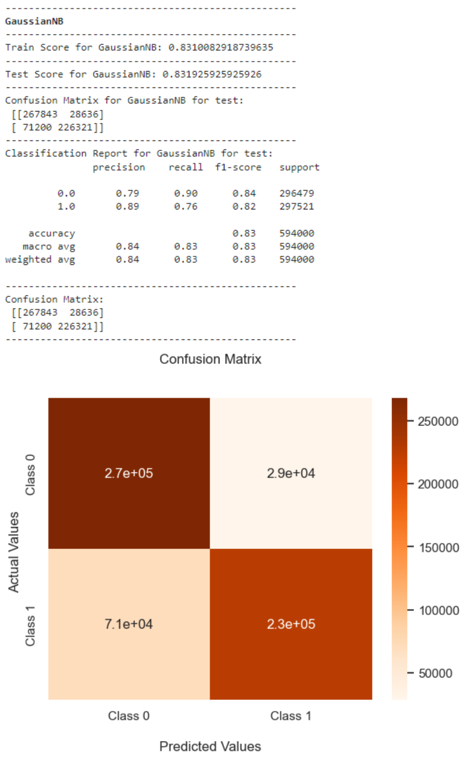
```
--------------------------------------------------
DecisionTreeClassifier
--------------------------------------------------
Train Score for DecisionTreeClassifier: 0.8509618573797678
--------------------------------------------------
Test Score for DecisionTreeClassifier: 0.8517508417508417
--------------------------------------------------
Confusion Matrix for DecisionTreeClassifier for test:
 [[289373   7106]
 [ 80954 216567]]
--------------------------------------------------
Classification Report for DecisionTreeClassifier for test:
               precision    recall  f1-score   support

         0.0       0.78      0.98      0.87    296479
         1.0       0.97      0.73      0.83    297521

    accuracy                           0.85    594000
   macro avg       0.87      0.85      0.85    594000
weighted avg       0.87      0.85      0.85    594000

--------------------------------------------------
Confusion Matrix:
 [[289373   7106]
 [ 80954 216567]]
--------------------------------------------------
```

Fig. 3. Decision Tree



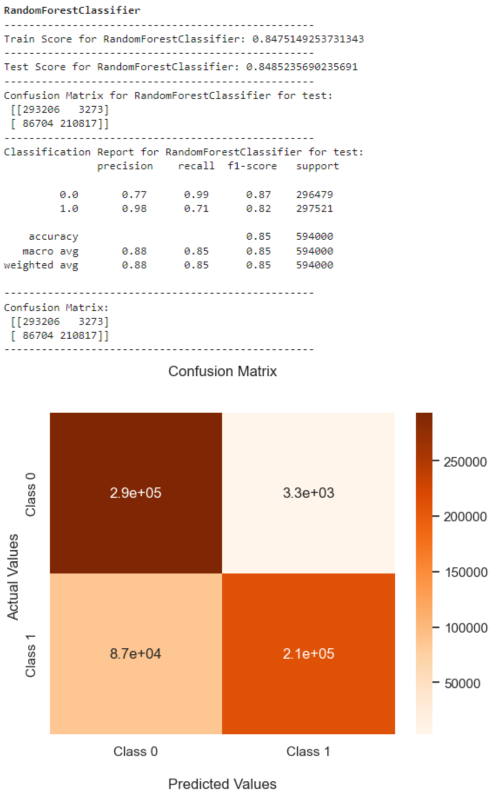
```
--------------------------------------------------
GaussianNB
--------------------------------------------------
Train Score for GaussianNB: 0.8310082918739635
--------------------------------------------------
Test Score for GaussianNB: 0.831925925925926
--------------------------------------------------
Confusion Matrix for GaussianNB for test:
 [[267843  28636]
 [ 71200 226321]]
--------------------------------------------------
Classification Report for GaussianNB for test:
               precision    recall  f1-score   support

         0.0       0.79      0.90      0.84    296479
         1.0       0.89      0.76      0.82    297521

    accuracy                           0.83    594000
   macro avg       0.84      0.83      0.83    594000
weighted avg       0.84      0.83      0.83    594000

--------------------------------------------------
Confusion Matrix:
 [[267843  28636]
 [ 71200 226321]]
--------------------------------------------------
```

Fig. 4. GaussianNB



```
RandomForestClassifier
--------------------------------------------------
Train Score for RandomForestClassifier: 0.8475149253731343
--------------------------------------------------
Test Score for RandomForestClassifier: 0.8485235690235691
--------------------------------------------------
Confusion Matrix for RandomForestClassifier for test:
 [[293206   3273]
 [ 86704 210817]]
--------------------------------------------------
Classification Report for RandomForestClassifier for test:
               precision    recall  f1-score   support

         0.0       0.77      0.99      0.87    296479
         1.0       0.98      0.71      0.82    297521

    accuracy                           0.85    594000
   macro avg       0.88      0.85      0.85    594000
weighted avg       0.88      0.85      0.85    594000

--------------------------------------------------
Confusion Matrix:
 [[293206   3273]
 [ 86704 210817]]
--------------------------------------------------
```

Fig. 5. RandomForestClassifier

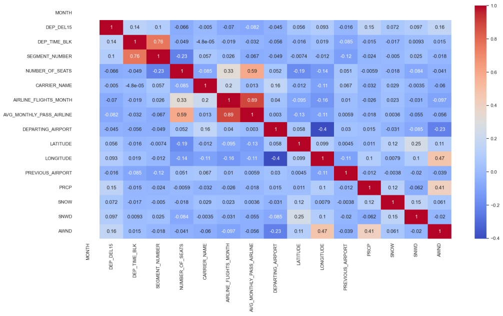


Fig. 6. Correlation Matrix

REFERENCES

[1] 2019 Airline Delays w/Weather and Airport Detail https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations

[2] Federal Aviation Administration (FAA). (2020). 2019 U.S. Aviation Statistic Annual Report. Washington, D.C.: FAA.

[3] National Center for Excellence in Aviation Operations (NCEAO). (2019). The Causes of Flight Delays and Cancellations: A Literature Review. Oklahoma City, OK: NCEAC.

[4] Smith, A. (2019). The Impact of Airline Delays on Passengers. New York, NY: The Travelers Bureau.

VIII. YOUTUBE VIDEO LINKI

Video Linki: https://youtu.be/NybScJYzdHs