A Synopsis Report

On

Flight Delay Predictor

For

Predictive Analytics

By

Yash Varshney - 500110167

Tanay Garg – 500108691

Under the guidance of

Dr. Achala Shakya

University of Petroleum and Energy Studies

Dehradun-India

Table of Contents

1. Abstract……2

2. Introduction……..2

3. Problem Statement……2

4. Objectives……3

5. Methodology……..3

6. Challenges faced……6

7. Results……6

8.Conclusion……6

10.References……7

Abstract:
The increasing need for reliable and accurate flight delay predictions has become crucial in the aviation industry, where unanticipated delays affect both passenger satisfaction and airline efficiency. This project focuses on developing a predictive model to determine flight delays based on various factors such as departure time, weather conditions, day of the week, and geographical data. Using machine learning techniques, particularly Gradient Boosting Classifier, the model leverages these features to predict delays with high accuracy. The goal is to create a system that provides real-time delay predictions to optimize scheduling, improve customer experience, and enhance operational efficiency. The project employs Indian geographical data to ensure relevance, with potential applications for airlines, airports, and passengers alike. The model's performance is evaluated through accuracy, precision, recall, and F1-score, demonstrating effective prediction results across various scenarios.

1. Introduction:
Flight delays cause inconvenience for passengers and significant financial losses for airlines. Predicting these delays with accuracy is a critical need in the aviation industry. This project aims to develop a machine learning-based flight delay predictor that can classify and estimate delays based on a variety of features, including time, origin and destination, and other contextual factors. By understanding the likelihood of delays, airlines can proactively manage schedules and improve the passenger experience.

2. Problem Statement:
Flight delays are affected by a range of factors, including weather conditions, airport congestion, and operational issues. Predicting delays is challenging because these factors are dynamic and interact in complex ways. Traditional approaches have struggled with accuracy, especially in real-time prediction, due to the wide variability in delay causes and limited data representation. The primary challenge is to develop a robust system that can handle this complexity and make accurate delay predictions in real-time scenarios, taking into account both scheduled and real-time data.

3. Objective:
The objective of this project is to create a model that can classify flights into different delay categories or predict the delay duration with high accuracy. The key delay categories include:

- **On-time**
- **15-30 minute delay**
- **30-60 minute delay**
- **60+ minute delay**

3.1 Applications of the model:

1. **Airline Operations**: Improve scheduling and operational efficiency by anticipating delays.
2. **Passenger Experience**: Enable passengers to make informed decisions based on predicted delays.
3. **Airport Management**: Optimize resource allocation at airports by predicting high-congestion periods due to delayed flights.

4. Methodology:

The project consists of several stages: data collection, preprocessing, model training, evaluation, and deployment.

4.1 Data Collection and Preprocessing:

- **Dataset**: Historical flight data from Indian airports, including variables like departure time, origin and destination airports, day of the week, and weather conditions.

4.2 Preprocessing:

- **Missing Values**: Handle missing or incomplete data entries.
- **Normalization**: Normalize numerical features for consistent processing.
- **Feature Engineering**: Create features such as day of the week, departure hour, and weather conditions, as these are crucial for capturing patterns related to delays.

4.3 Model Development:

- **Gradient Boosting Classifier**: Gradient boosting is selected for its ability to handle structured tabular data with strong predictive performance.

4.4 Model Architecture:

- Input Layer: Features such as origin, destination, departure time, and weather conditions.
- Boosting Stages: Series of weak learners combined to minimize prediction error.
- Output Layer: Softmax or regression layer to output delay classification or delay time estimation.

4.5 Model Training and Evaluation:

- **Training and Validation Split**: Split the data into training and validation sets to evaluate model performance.
- **Loss Function**: Categorical cross-entropy for classification and Mean Squared Error (MSE) for regression.
- **Optimization**: Hyperparameter tuning to find the optimal configuration.
- **Metrics**: Evaluate model performance using accuracy, precision, recall, F1-score, and Mean Absolute Error (MAE).
- **Cross-validation**: Use K-fold cross-validation to ensure robustness across different data splits.

5. Deployment :

- Deploy the model as a web application where users can input flight details and receive real-time delay predictions.
- **Technologies**: Use Flask or Django for web deployment or a simple GUI with Tkinter for desktop applications.

6. Technologies and Tools:

- **Programming Language**: Python
- **Libraries**:
  - **Scikit-Learn**: For implementing the Gradient Boosting Classifier model.
  - **Pandas & NumPy**: For data manipulation and analysis.
  - **Matplotlib & Seaborn**: For visualizations, such as feature importance and prediction accuracy.
- **Development Environment**: Jupyter Notebook or any Python IDE like PyCharm.
- **Version Control**: Git and GitHub for version control and collaboration.

7. Challenges Faced:

1. **Feature Engineering**: Identifying the most impactful features for delay prediction required iterative experimentation.
2. **Data Imbalance**: Delays in some categories were underrepresented. Resampling techniques, such as oversampling, were used to address this issue.
3. **Real-Time Processing**: Processing real-time data efficiently was challenging, with potential improvements using optimized deployment.
4. **Hyperparameter Tuning**: Finding the best parameters for the Gradient Boosting Classifier to achieve optimal performance took considerable time.

8. Results:

- **Accuracy**: The model achieved an accuracy of approximately 80%, accurately classifying most delay instances.
- **Precision/Recall**: High precision and recall scores were obtained for on-time and minor delay classes, though longer delays had slightly lower performance.
- **Confusion Matrix**: Analysis of the confusion matrix showed occasional misclassification between short and moderate delays.

9. Conclusion:

The Flight Delay Predictor project demonstrates the utility of machine learning, particularly Gradient Boosting, in forecasting delays. By accurately predicting delays, this model supports airline operations, passenger planning, and airport management. Although the model performed well, further improvements can be made to enhance prediction accuracy, particularly for longer delays.

References:

[1] Indian Flight Dataset : https://github.com/OludolapoAnalyst/Indian_Flight_Data

[2] GeoPandas Documentation : https://geopandas.org/en/stable/docs.html

[3] India Shapefile : https://www.indiaremotesensing.com/2017/01/download-india-shapefile-with-official.html