Predicting Flight Delays Using Data Mining Algorithms on US Flight Data

Aishwarya Devi Akila Pandian
Purdue University
West Lafayette, Indiana, USA
aakilapa@purdue.edu

Sai Rahul Reddy Kondlapudi
Purdue University
West Lafayette, Indiana, USA
skondlap@purdue.edu

Shrujan S Katama
Purdue University
West Lafayette, Indiana, USA
skatama@purdue.edu

Vigneshwaran Dharmalingam
Purdue University
West Lafayette, Indiana, USA
vdharmal@purdue.edu

Abstract

Flight delays remain one of the most prevalent issues in modern aviation, costing both airlines and passengers time and money. Our project aims to predict flight delays using data mining algorithms applied to the Kaggle dataset, "Flight Delay Exploratory Data Analysis." By leveraging machine learning techniques, we aim to build robust models to predict delays and analyze the key contributing factors, such as weather, flight duration, and airline performance. We will compare three popular data mining algorithms: Random Forest, Gradient Boosting (XGBoost), and Support Vector Machines (SVM), and evaluate their performance in terms of accuracy, computational efficiency, and interpretability. The insights gained from this project can be useful for airlines to optimize their operations and improve customer satisfaction.

1 Introduction

Flight delays are an ongoing challenge for the aviation industry, affecting millions of travelers worldwide and causing significant economic losses. The causes of flight delays are often multifaceted, ranging from adverse weather conditions to operational inefficiencies. This complexity makes flight delay prediction a suitable problem for data mining techniques, as they can capture non-linear relationships and interactions between variables that simpler methods might miss.

Our project proposes to predict flight delays using real-world data from the Kaggle dataset, which includes detailed information such as flight times, distances, and weather conditions. We plan to apply and compare three data mining algorithms—Random Forest, XGBoost, and Support Vector Machines (SVM)—to identify patterns in the data and predict delays. By conducting a thorough analysis of the dataset and evaluating algorithm performance, we aim to uncover insights that could assist airlines in minimizing delays and improving operational efficiency.

2 Team Composition

Our team consists of four members:

- Team Member 1: [Aishwarya Devi Akila Pandian] – Data preprocessing and Exploratory Data Analysis (EDA)
- Team Member 2: [Sai Rahul Reddy Kondlapudi] – Algorithm implementation and hyperparameter tuning
- Team Member 3: [Shrujan S Katama] – Model evaluation and comparison
- Team Member 4: [Vigneshwaran Dharmalingam] – Literature review and report writing

Each team member will focus on different aspects of the project to ensure a collaborative and effective approach to building predictive models.

3 Proposed Work

3.1 Dataset and Problem Description

The dataset used for this project is sourced from Kaggle and contains a comprehensive set of features that can impact flight delays, including scheduled departure and arrival times, actual flight durations, weather data, airline codes, and airport information. The challenge lies in handling the missing data, potential noise, and large variability in the dataset.

Our project will focus on transforming this raw data into a structured format suitable for training machine learning models. We will perform extensive feature engineering, including the extraction of relevant temporal features (e.g., time of day, day of the week), weather conditions, and airline-specific factors that might influence delays.

The primary problem we are addressing is the binary classification of flights into two categories: delayed and not delayed. We will formulate this as a supervised learning task where the delay threshold is set at 15 minutes, a commonly used metric in airline reporting.

3.2 Planned Activities

We plan to execute the project in several stages:

- **Literature Survey**: We will begin with a review of related work in flight delay prediction, focusing on both traditional statistical methods and modern machine learning approaches. This will help us refine our methodology and set appropriate benchmarks for our algorithms.
- **Data Collection & Exploration**: After obtaining the dataset, we will conduct Exploratory Data Analysis (EDA) to understand its structure and identify potential correlations between variables. We will handle missing data, outliers, and apply necessary data transformations such as one-hot encoding for categorical variables (e.g., airlines).

- **Algorithm Selection & Design**: The core of our project will be the implementation of three data mining algorithms:
  (1) **Random Forest**: A widely-used ensemble learning method that handles large datasets and provides insights into feature importance.
  (2) **XGBoost**: A boosting algorithm known for its high predictive accuracy and efficiency in handling structured data.
  (3) **SVM (Support Vector Machines)**: A powerful classification technique that works well with high-dimensional data and has strong theoretical foundations.

  Hyperparameter tuning will be conducted for each model to optimize their performance.
- **Model Evaluation**: We will evaluate the models based on key performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Computational efficiency will also be considered, especially given the large size of the dataset.
- **Insights & Analysis**: After training and evaluating the models, we will analyze feature importance scores (e.g., from Random Forest and XGBoost) to identify the key factors contributing to flight delays. This analysis could offer valuable insights for airlines to improve their scheduling and operational processes.

3.3 Evaluation Plan

Our evaluation process will include both quantitative and qualitative metrics:

- **Performance Metrics**: We will assess the predictive accuracy of the models using precision, recall, F1-score, and AUC-ROC. We will also perform cross-validation to ensure the robustness of our models.
- **Feature Importance**: Using models like Random Forest and XGBoost, we will extract feature importance scores to understand which variables (e.g., weather, departure time) most strongly impact delays.
- **Visualization**: Results will be presented using visual tools like confusion matrices and feature importance bar charts to aid in interpretability. Comparative graphs of model performance will help us determine the best-performing model.

4 Project Timeline

The project timeline is organized into several phases, starting from the literature review and ending with the final report and presentation. Each phase will ensure steady progress and thorough analysis throughout the course of the project.

Activity	Timeline
Literature Survey	Sep 23 - Sep 30, 2024
Data Collection & Exploration	Oct 1 - Oct 14, 2024
Algorithm Selection & Design	Oct 15 - Oct 31, 2024
Midterm Report	Oct 23, 2024
Model Evaluation & Comparisons	Nov 1 - Nov 7, 2024
Feature Importance & Insights	Nov 8 - Nov 19, 2024
Final Report Preparation	Nov 20 - Dec 8, 2024
Final Presentation	Nov 26 - Dec 5, 2024

Table 1: Project Timeline

5 Conclusion

This project will provide hands-on experience in the application of data mining algorithms to a complex, real-world problem. By predicting flight delays and analyzing the underlying causes, we hope to offer actionable insights that could assist airlines in improving their operational efficiency. We believe that comparing Random Forest, XGBoost, and SVM will highlight the strengths and weaknesses of each model in terms of accuracy, interpretability, and computational cost, providing a comprehensive analysis of their effectiveness in this domain.

6 References

- Dataset: https://www.kaggle.com/code/robikscube/flight-delay-exploratory-data-analysis-twitch/notebook