**Data Analysis Report: Predicting Flight Delays – An In-Depth Analysis**

**Executive Summary**

This report addresses the challenge of predicting flight delays of 50 minutes or more by analysing flight data. Emphasis is placed on key factors including the airline, day of the week, and airport of origin. Utilizing advanced machine learning techniques, specifically the XGBoost algorithm, we developed and optimized a predictive model aimed at forecasting these delays. The findings of this analysis underscore the value of data-driven decision-making within the airline industry, presenting opportunities to enhance operational efficiency, improve customer experience, and mitigate costs.

**1. Situation**

Flight delays represent a significant challenge within the airline industry, leading to disruptions for passengers, increased operational expenses for airlines, and a detrimental effect on overall customer satisfaction. The ability to accurately predict such delays in advance is essential for optimizing resource management, enhancing flight scheduling, and improving the travel experience. This analysis aims to address this issue by identifying the factors that contribute to flight delays and developing a predictive model capable of forecasting whether a flight will be delayed by 50 minutes or more.

**2. Task**

The primary objectives of this initiative were to:

1. Conduct a thorough analysis of the data to identify patterns and trends that contribute to flight delays.
2. Develop a predictive model utilizing machine learning techniques to forecast significant flight delays (those exceeding 50 minutes).
3. Enhance the model's accuracy and performance through optimization methods, including oversampling techniques such as SMOTE and hyperparameter tuning.
4. Present actionable insights that could assist airlines in optimizing their operations, mitigating delays, and enhancing overall customer satisfaction.

**3. Action**

**3.1 Data Preprocessing and Exploration**

We commenced our analysis by thoroughly examining the dataset, which comprised various variables including departure time, airline information, origin and destination airports, as well as delay metrics. Our focus centred on several key variables, specifically:

* Day\_of\_Week: The designated day of the week for the flight.
* Date: The specific date of the flight is reformatted into month and day for enhanced analytical clarity.
* Dep\_Time: The scheduled departure time.
* Airline: The airline responsible for operating the flight.
* Origin: The airport code representing the departure location.
* Carrier\_Delay: The delay attributable to the airline's operations.

**3.2 Exploratory Data Analysis (EDA)**

\*\*Delays by Airline\*\*: We calculated the average delay percentages across different airlines, revealing that American Airlines Inc. experienced the highest delay rate at 82.33%, while Frontier Airlines Inc. recorded the lowest at 7.73%.

\*\*Delays by Day of the Week\*\*: Our analysis indicated variability in delays throughout the week, with Saturday (Day 6) illustrating the highest delay percentage at 58.8%, in contrast to Sunday (Day 7), which exhibited the lowest at 34.0%.

\*\*Delays by Origin Airport\*\*: Certain airports demonstrated significant delay rates, such as ABI with an extreme 99.91%, while ATL was noted for a more moderate delay rate of 34.76%.

### 3.3 Feature Engineering and Model Preparation

Data preparation involved converting the Date column into distinct month and day variables, followed by the elimination of the original Date column. We identified categorical variables including Airline, Origin, and Destination, which were subsequently encoded for incorporation into the machine learning model.

### 3.4 Model Building and Evaluation

\*\*Initial Model (XGBoost)\*\*: We employed the XGBoost model for preliminary predictions, achieving a baseline accuracy of 89.62% and an AUC score of 0.6726. While these results indicated a satisfactory performance level, there remained opportunities for enhancement, particularly in the accurate prediction of delayed flights.

\*\*Handling Class Imbalance (SMOTE)\*\*: Due to the inherent imbalance within the dataset—characterised by a higher frequency of non-delayed flights compared to delayed ones—we applied the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic samples of the minority class. This intervention was instrumental in balancing the dataset and improving overall model performance.

\*\*Hyperparameter Tuning\*\*: We conducted a systematic hyperparameter tuning process to optimize the model’s performance. The most effective parameters identified were:

- \*\*n\_estimators\*\*: 250

- \*\*subsample\*\*: 0.6

The implementation of these optimizations resulted in a significant enhancement of the AUC score, which increased to 0.961, thereby markedly improving the model’s efficacy in predicting flight delays.

**4. Results**

**4.1 Model Performance**

Upon finalizing the model, the optimized XGBoost model achieved an accuracy rate of 89.23%. However, the Area Under the Curve (AUC) score was marginally lower at 0.6665, suggesting that there remains potential for enhancing the model's proficiency in accurately detecting delayed flights. The following confusion matrix illustrates the performance metrics:

True Negatives (TN): 128,520

False Positives (FP): 2,940

False Negatives (FN): 12,720

True Positives (TP): 1,186

Although the accuracy is commendable, the model would benefit from additional refinements, particularly in the reduction of false negatives, which represent missed delay predictions.

**5. Reflection**

**5.1 Insights**

The impact on airlines varies significantly, with certain carriers, such as American Airlines Inc., exhibiting notably higher percentages of delayed flights. Addressing operational inefficiencies within these airlines may contribute to a reduction in overall delays. Furthermore, the day of the week has a pronounced effect on flight delays, with an observable increase during weekends, particularly on Saturdays. It may be prudent for airlines to optimize staffing and resource allocation during these peak periods to mitigate delays.

Additionally, specific airports, such as ABI, demonstrate exceptionally high delay rates, suggesting that these locations confront unique challenges, including logistical issues or inadequate resources. Implementing targeted improvements at these airports could lead to a substantial decrease in delays.

**5.2 Recommendations**

Airlines should prioritize enhancements in operational efficiency, particularly during weekends and at airports characterized by high delay rates. Moreover, it is essential to refine the predictive model further, potentially by incorporating additional features or real-time data to enhance its accuracy. By leveraging this analytical framework, airlines can optimize flight scheduling, improve resource allocation, and proactively manage delays, thereby enhancing overall customer satisfaction.

6. Conclusion

This project demonstrates the effectiveness of using data-driven insights to predict flight delays and identify factors that contribute to operational inefficiencies. The XGBoost model, with further tuning, can be an essential tool for airlines to predict delays, improve operational strategies, and enhance the travel experience. This end-to-end analysis, using machine learning techniques and data visualization, offers actionable insights that can lead to significant improvements in the airline industry.

7. Next Steps

Explore additional features such as weather conditions, flight distance, and aircraft maintenance issues to enhance the model.

Implement real-time data inputs to predict delays for upcoming flights and dynamically adjust operations.

Collaborate with airlines to implement this model in operational decision-making to optimize flight schedules and improve customer satisfaction.