Aviation Anomaly Detection of Delayed Civilian Flight Departures

Akileiswaran Samuthiran (A0111578E), Chung YeonJong (A0280412R), Jared Cheang (A0222660N), Lim Eng Xiong (A0053475A), Lim Yi Kai (A0280507H)

Abstract

Departure delays significantly impact both airline operations and passenger itineraries, with inefficient claim processes often exacerbating such stressful situations. This project explores machine learning (ML) approaches for predicting flight departure delays. By reducing the unpredictability of flight disruptions, a travel insurance company can design higher coverage plans and scale their resources to enhance customer satisfaction. Both a continuation and pivot of the initial proposal which aims at enhancing flight anomaly detection for Air Traffic Controllers, this study retains the core competency in anomaly detection and extends its impact to consumer-facing applications. A 2023 US civilian flight dataset is used for the scope of this project. The overall approach involves exploratory data analysis (EDA), feature selection and engineering to derive the most suitable features. An AUC score of 0.74 was achieved using an XGBoost classifier, after evaluating multiple ML models, demonstrating the feasibility and potential of leveraging ML in the aviation industry.

1. Introduction

1.1 Background

Flight delays are an increasing concern in the aviation industry, significantly affecting airline operations and passenger itineraries. In 2023, various US airports recorded significant delays, with budget airlines and some major airports like Fort Lauderdale-Hollywood International showing a higher incidence of delays due to factors such as NAS-related issues and late arriving aircraft[1]. This rise in delays poses challenges not only for air traffic management but also impacts passengers profoundly, leading to missed connections, disrupted travel plans, and often, substantial unplanned expenses.

Amid these operational disruptions, travel insurance emerges as a critical tool for mitigating the financial risks associated with flight delays. Travel insurance policies typically cover emergency medical expenses, trip cancellations, delays, and lost or stolen baggage, providing essential protection in the unpredictable realm of travel. Travelers are advised to purchase insurance that matches the level of risk they are comfortable with, often choosing between basic coverage plans for less frequent travelers and more comprehensive multi-trip policies for those who travel regularly.

The process of filing claims for delays under travel insurance policies is also critical. It generally involves timely notification to the insurer, submission of relevant documentation such as proof of delay, and sometimes direct communication between the insurer and the service providers like airlines or hospitals. This process underscores the importance of understanding the terms and coverage limits of insurance policies to effectively manage and mitigate the impacts of travel disruptions.

By exploring machine learning methods, this study approaches to predict flight delays, aims to enhance the proactive capabilities of both air traffic controllers and insurance companies, ultimately leading to more resilient and responsive operations in the face of increasing anomalies in flight schedules.

1.2 Problem Statement

In travel insurance, companies face significant challenges in managing customer satisfaction and operational efficiency. Flight delays frequently disrupt travel plans, leading to an increase in insurance claims, which travel insurance providers must handle promptly. Customers expect quick responses and resolutions to their service requests, especially during high-volume periods triggered by widespread travel disruptions, such as those experienced during the COVID-19 pandemic. Furthermore, insurance providers struggle with policy pricing and risk management. They need to devise strategies that not only capture the correct demographic with competitive but also profitable premium rates but ensure that customers perceive the value, justifying higher premiums in exchange for comprehensive coverage and reliable service. These challenges highlight the need for improved customer service protocols, innovative risk assessment models, and

[1] https://travelfreak.com/airline-delay-statistics/

strategic premium pricing to enhance customer satisfaction and company profitability in the travel insurance sector.

1.3 Objectives

This project's main objective is to develop a predictive model that accurately forecasts flight delays exceeding 60 minutes. This specific target threshold is chosen because a delay of over an hour significantly disrupts passenger itineraries, making it a critical point at which travelers most often seek compensation and support from their travel insurance. By focusing on delays that surpass this 60-minute mark, the selected model addresses the most impactful disruptions that affect passengers, aligning closely with the needs and expectations of both travelers and insurance providers.

By providing a service that forecasts with high reliability, the model not only supports the operational efficiency of travel insurance providers but also builds trust among customers. This trust is underpinned by a strong record of accomplishment and a consistently good hit rate, which reassures customers of the model's efficacy and the soundness of basing their travel decisions on its recommendations. This project seeks to refine the prediction of flight delays of more than 60 minutes, enhancing the value proposition of travel insurance in managing travel uncertainties.

2. Data Discovery

2.1 Data Description

The dataset used for this study is sourced from Kaggle, it provides civil aviation information in the US from Jan 2023 to Dec 2023, supplemented by daily meteorological conditions for the induvial airports and their respective regions. While the dataset focuses on US flights, it is highly representative of broader trends in the aviation industry due to the following reasons:

- Market Share: The US had the largest commercial air travel market in 2021 (Statista, 2023) with over 666 million passengers. It has a diverse and extensive air transport network, good range of scale and operational capabilities of its many airports to represent conditions around the world.

- Weather Variations: Flight delays caused by adverse weather conditions are a global problem, seasonal and meteorological patterns affecting US flights can also be commonly experienced worldwide.

The origin of flight records is the Bureau of Transportation Statistics (BTS) while the National Weather Service (NWS) allowed extraction of publicly available meteorological conditions, all consolidated into the Kaggle dataset. *Table 1 and Table 2* below describe the two sets of data features.

Table 1. US Flights in 2023

FEATURE	DESCRIPTION
FlightDate	Departure date in yyyy-mm-dd
Day_Of_Week	1-7, where 1 = Monday
Airline	Names of Airline companies
Tail_Number	Unique Tail Number for aircraft
Dep_Airport	Abbreviated Departure Airport
Dep_CityName	Departure city name
DepTime_Label	Departure period of the day
Dep_Delay	Departure delay in minutes
Dep_Delay_Tag	Departure delay indicator > 5min
Departure_ Delay_Type	Departure delay duration category
Arr_Airport	Abbreviated Arrival Airport
Arr_CityName	Arrival city name
Arr_Delay	Arrival delay in minutes
Arr_Delay_Type	Arrival delay duration category
Flight_Duration	Flight duration in minutes
Distance_Type	Distance category
Manufacturer	Aircraft manufacturer name
Model	Aircraft model
Aircraft_age	Aircraft age
Delay_{X}	Remaining columns where X is in [Carrier, Weather, NAS, Security, LastAircraft].

Table 2. Weather and Meteorological Records

FEATURE	DESCRIPTION
time	Date (yyyy-mm-dd)
tavg	Average Temperature (°C)
tmin	Minimum Temperature (°C)
tmax	Maximum Temperature (°C)
prcp	Total Precipitation (mm)
snow	Snow Depth

wdir	Wind (From) Direction (Degrees)
wspd	Average Wind Speed (km/h)
pres	Sea-Level Air Pressure (hPa)
airport_id	Unique ID airport

It is recognized that the dataset does possess some key limitations. Firstly, it contains only domestic flight records within the US. International flights, which often involve longer travel times and more complex itineraries, are not included. This has a greater impact on passengers' travel plans and higher financial stakes. Thus, an assumption was made that this domestic flight data should provide contain similar patterns for identifying departure delays. It is also noted that the dataset covers only one year that could limit effectiveness when training the ML models to capture seasonal patterns. With only one year's worth of data, it is unable to establish any relationship regarding flight schedules, frequency and meteorological records with seasons, vacation, or holidays periods.

2.2 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) focused on two delay metrics: overall delay and severe delay. A new binary feature on severe delays was created to identify delays exceeding 60 minutes. To gain a broader understanding of delay patterns, including minor delays, an additional analysis considered delays greater than 0 minutes. The results showed that 6.88% of flights experienced delays exceeding 60 minutes, while 37.9% encountered some form of delay (greater than 0 minutes).

In preparation for model training, features unavailable during prediction were excluded. This included all delay-related factors. However, the feature on flight duration was retained as most flights have estimated flight times available before takeoff. Finally, the data was split for training and testing purposes. Bearing in mind the whole dataset spans an entire year of 2023 from January to December, the most recent months, November, and December, were thus used for testing, while the remaining data consisting of the earlier months in the year formed the training set.

The EDA revealed several key insights about factors potentially influencing any delays and severe delays as follows.

- Day of Week: The distribution of flights across weekdays was relatively even, suggesting no significant impact on delays.

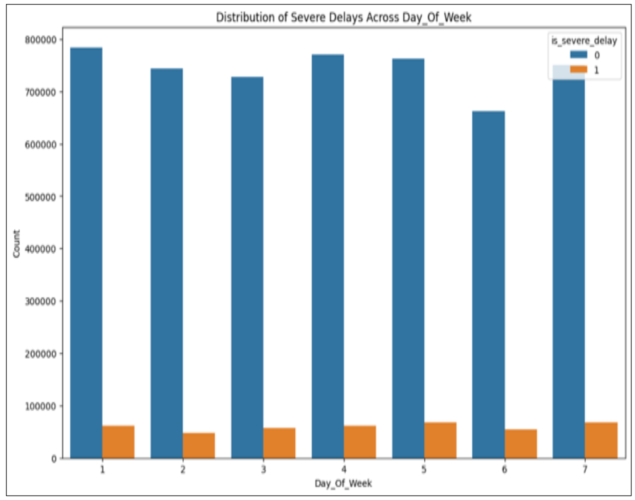


Figure 1. Distribution of Severe Delays Across "Day_Of_Week"

- Airline Carrier: The distribution of airline carriers was skewed. To address dimensionality, the top K carriers were identified based on their contribution to severe delays. By analyzing this relationship later, a value of K=7 was chosen, capturing 63% of the population. Carriers within this group were assigned individual labels, while the remaining carriers were grouped as "Others".

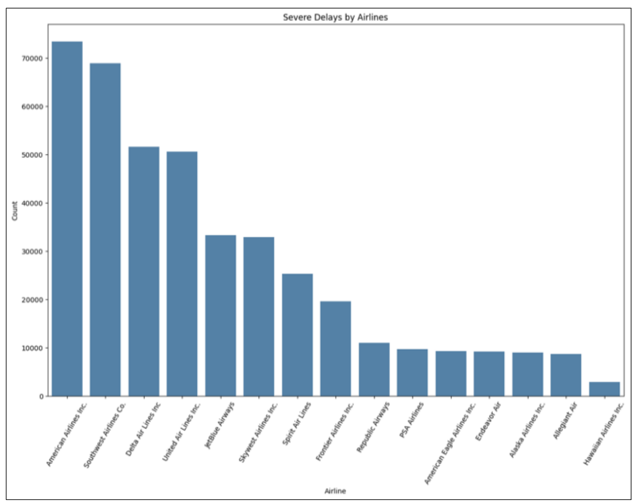


Figure 2. Severe Delays by Airline.

- Departure Airport: Like airline carriers, departure airports exhibited a skewed distribution. Here too, the top K airports were identified based on their contribution to severe delays. Analyzing this relationship led to selecting K=50, capturing a significant portion of the data while reducing dimensionality of the dataset.

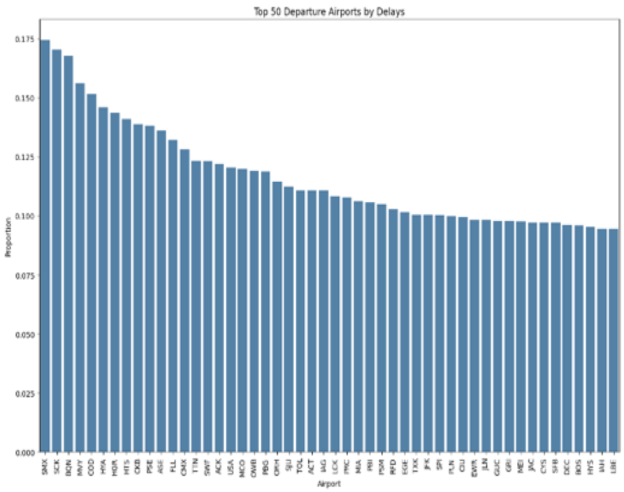
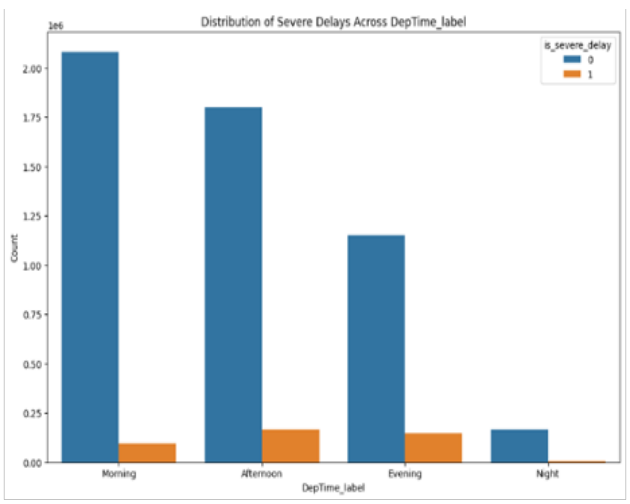


Figure 3. Top 50 Departure Airports by Delays.

- Departure Time: Flight departures were concentrated in the afternoons and evenings,



regardless of whether the flight experienced a delay or a severe delay.

Figure 4. Distribution of Severe Delays Across 'DepTime_label'

- Model Type: The DA40 aircraft model exhibited a disproportionately high delay proportion compared to other models. A dedicated feature was created to identify DA40 flights for further analysis.

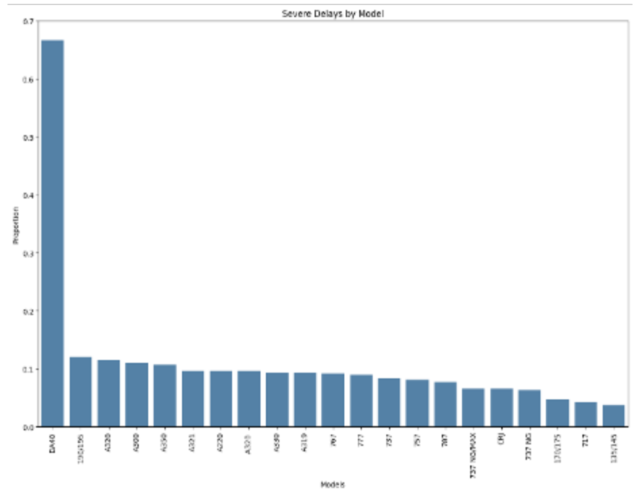


Figure 5. Severe Delays by Airline Model

- Aircraft Age: No obvious linear correlation was observed.

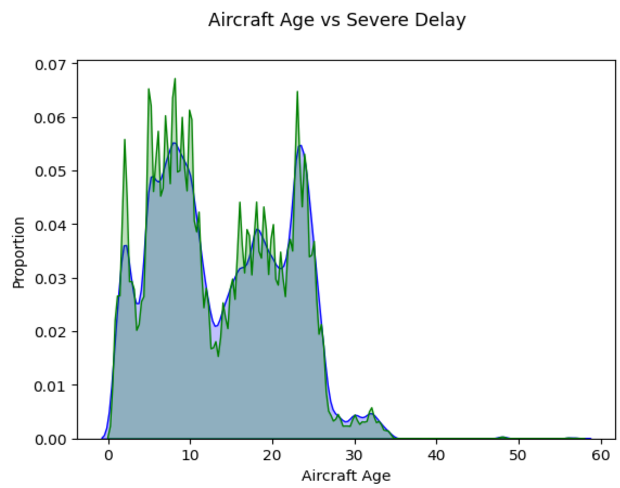


Figure 6. Aircraft Age vs Severe Delay

- Feature Correlations: Correlations between features were examined. As expected, a correlation was observed between the occurrence of any delay and the occurrence of a severe delay. Otherwise, most other features yield a weak correlation to either a delay or severe delays (<0.1). There was no direct correlation between the model being a DA40 and the likelihood of a delay.

Figure 7. Correlation heatmap with Any and Severe Delays

3. Methodology

3.1 Summary of Approach and Assumptions

In this study, a comprehensive pipeline was developed for building and validating a machine learning model aimed at predicting flight delays as follows:

- Data Collection: Gathering comprehensive civil aviation and meteorological data.
- Data Preprocessing: Cleaning and preparing the data for analysis. Since the main objective of this project is to accurately forecast flight delays, features within the dataset which were collected post-delay were excluded. For example, post-delay data such as arrival delay data was excluded since it does not influence the departure delay of the corresponding flight.
- Feature Engineering: Transforming the data into useful features that enhance model prediction performance, focusing on airline and weather datapoints.
- Model Development: Constructing the model pipelines and experimenting across various models for delay prediction. Here, traditional and tree-based models were explored, where the latter category was expanded on due to their strengths in capturing non-linear relationships and ability to leverage on ensemble techniques.
- Model Comparison and Selection: Training the various models then evaluating against a common test set for model comparison.
- Model Explainability: Applying Explainable AI techniques on explored models to validate their reasoning; this ensures the analysis aligns with existing understanding of what causes delays and identifies any unexpected factors influencing the predictions.

3.2 Data Preprocessing and Feature Engineering

The feature engineering process aimed to transform raw data into a format suitable for model training. Here is a breakdown of the key steps:

- Airline Carrier Categorization: Carriers were categorized into two groups: a "Top K" group consisting of the K carriers with the highest contribution to severe delays, and an "Others" group encompassing the remaining carriers.
- Model Type Identification: A new feature was created to specifically identify flights operated by the DA40 aircraft model.
- Departure Airport Delay Propensity: A delay probability score was assigned to each departure airport based on historical data of severe delays at that airport and the airline carrier operating the flight. This score leverages past delay patterns to potentially enhance prediction accuracy.
- Weather Feature Creation: New features were introduced to capture weather conditions potentially impacting delays. These features included freezing temperatures, rain occurrence, and snow occurrence.
- Wind Limit Feature: A new feature was created to account for wind limitations during takeoff. This feature considered both dry and wet conditions based on airline guidelines, focusing on the departure scenario. Focusing on the Boeing 737-800 which is a widespread modern air carrier, existing guidelines suggest that maximum allowable crosswinds are around 33 knots for dry runways and are reduced to approximately 27 knots for wet runways[2].

[2] https://www.flightdeckfriend.com/ask-a-pilot/aircraft-maximum-wind-limits/

- Data Reduction: To minimize potential bias and focus on broader patterns, several features were excluded from the training data. These included departure airport and city details, airline manufacturer, airline model, and month of travel.
- Column Encoding: Categorical features such as day of week, airline carrier, and departure time were encoded using one-hot encoding. This approach transforms categorical data into binary features, improving model interpretability.
- Ordinal Encoding: The Distance Type feature, which represents flight distance categories, was encoded using ordinal encoding. This technique acknowledges the natural order inherent in these distance categories (e.g., short-haul, medium-haul, long-haul).

3.3 Model Development

To ensure both efficiency and reproducibility in model development, the 'sklearn.pipeline' library was utilized to construct pipelines for preprocessing and model training. These pipelines transform categorical input features using either one-hot encoding or ordinal encoding, followed by feature normalization which helps to prevent the model from disproportionately weighing features based on their scale. Each pipeline culminates in fitting a model that predicts the probability of a flight being severely delayed.

The initial set of models explored includes:

- Logistic Regression: Employed as a base model, it assumes linear relationships between input features and the target variable 'is_severe_delay.' To counteract class imbalance, the class_weight parameter is set to 'balanced,' and max_iter is configured to 1000 to ensure model convergence.
- Decision Tree: This model addresses non-linear relationships that logistic regression may miss. The max_depth was limited to 5 to avoid over-complexity.
- Random Forest Classifier: As an ensemble of decision trees, this model generally outperforms a single decision tree but requires more computational resources. The hyperparameters n_estimators are set to 100, max_features to 'sqrt', and max_depth to 5, to assess significant improvements over the decision tree.
- Gradient Boosting Classifier: Another ensemble technique, using similar settings for n_estimators and max_depth as the random forest.
- Extreme Gradient Boosting (XGBoost) Classifier: An enhancement of the gradient boosting framework, employing the same hyperparameters to leverage its advanced capabilities.

This selection of models covers a range of assumptions and complexities, from linear to more sophisticated ensemble methods, facilitating a comprehensive evaluation of predictive performance under varied modeling conditions.

3.4 Model Comparison and Selection

While most of the models demonstrate impressively high training and testing accuracies, with scores exceeding 90% and 95% respectively, these metrics do not necessarily reflect the true performance due to the significant class imbalance within the dataset. Specifically, with only 7% of flights categorized as severely delayed, a model predicting all flights as timely would still achieve an accuracy of 93%. To better evaluate the models, the AUC score was therefore adopted as the primary metric, which provides a more comprehensive summary of a classifier's performance across various decision thresholds.

Upon reviewing the AUC scores, all models displayed moderate discrimination capabilities, with training scores ranging between 0.68 to 0.72 and testing scores from 0.62 to 0.64. However, the precision analysis revealed that the decision tree and random forest models were predicting all instances as non-severe delays, evidenced by their zero precision and recall, thus failing to effectively identify severe delays.

In contrast, the XGBoost model showed more potential, achieving a training precision of 0.64 and a testing precision of 0.39.

Table 3. Training Evaluation Metrics

Model	AUC	Precision	Recall	Accuracy
Logistic Regression	0.6802	0.1215	0.6260	0.6375
Decision Tree	0.6647	0.0000	0.0000	0.9260
Random Forest	0.6784	0.0000	0.0000	0.9260
Gradient Boosting	0.7036	0.6737	0.0070	0.9263
XGBoost	0.7193	0.6401	0.0140	0.9265

Table 4. Testing Evaluation Metrics

Model	AUC	Precision	Recall	Accuracy
Logistic Regression	0.6368	0.07263	0.4728	0.7155
Decision Tree	0.6230	0.0000	0.0000	0.9566
Random Forest	0.6278	0.0000	0.0000	0.9566
Gradient Boosting	0.6435	0.3824	0.0003	0.9567
XGBoost	0.6446	0.3904	0.0023	0.9566

To enhance the performance of the XGBoost model, fine-tuning of its hyperparameters was then conducted through a grid search, focusing on 'n_estimators' with values [100, 150, 200], and 'max_depth' with values [5, 10, 15]. Additionally, Stratified K-Fold cross-validation with 5 splits was implemented to ensure a robust assessment of the model's performance and more reliable tuning outcomes.

The best performing XGBoost model (n_estimators = 150, max_depth = 10) yields a mean train roc auc of 0.7410 and a mean train precision of 0.611.

Table 5: Cross validation results on hyperparameters grid search for XGBoost model

#	0	1	2	3	4	5	6	7	8
max_depth	5	5	5	10	10	10	15	15	15
n_estimators	100	150	200	100	150	200	100	150	200
mean_train_precisior	0.6254	0.6302	0.6308	0.6237	0.6113	0.6012	0.5228	0.4889	0.4641
std_train_precision	0.0125	0.0087	0.0056	0.0044	0.0019	0.0035	0.0021	0.0017	0.0031
rank_train_precision	3	2	1	4	5	6	7	8	9
mean_train_recall	0.0152	0.0197	0.0233	0.0568	0.0625	0.0662	0.0857	0.0914	0.0963
std_train_recall	0.0004	0.0008	0.0006	0.0012	0.0009	0.0010	0.0008	0.0010	0.0007
rank_train_recall	9	8	7	6	5	4	3	2	1
mean_train_f1	0.0297	0.0383	0.0450	0.1042	0.1135	0.1193	0.1472	0.1540	0.1595
std_train_f1	0.0008	0.0015	0.0011	0.0019	0.0014	0.0016	0.0011	0.0015	0.0009
rank_train_f1	9	8	7	6	5	4	3	2	1
mean_train_balanced_accuracy	0.5072	0.5094	0.5111	0.5271	0.5297	0.5314	0.5397	0.5419	0.5437
std_train_balanced_accuracy	0.0002	0.0004	0.0003	0.0006	0.0004	0.0005	0.0004	0.0005	0.0003
rank_train_balanced_accuracy	9	8	7	6	5	4	3	2	1
mean_train_roc_auc	0.7150	0.7197	0.7232	0.7398	0.7410	0.7408	0.7294	0.7240	0.7190
std_train_roc_auc	0.0004	0.0005	0.0006	0.0005	0.0006	0.0006	0.0007	0.0010	0.0011
rank_train_roc_auc	9	7	6	3	1	2	4	5	8

3.5 Model Explainability

With feature importance analysis, factors that are consistent across 3 models are precipitation, departure time, and temperature; these make sense as direct influence on flight external factors and takeoff conditions; potential relationship to lighting, rain, snow, or ice conditions.

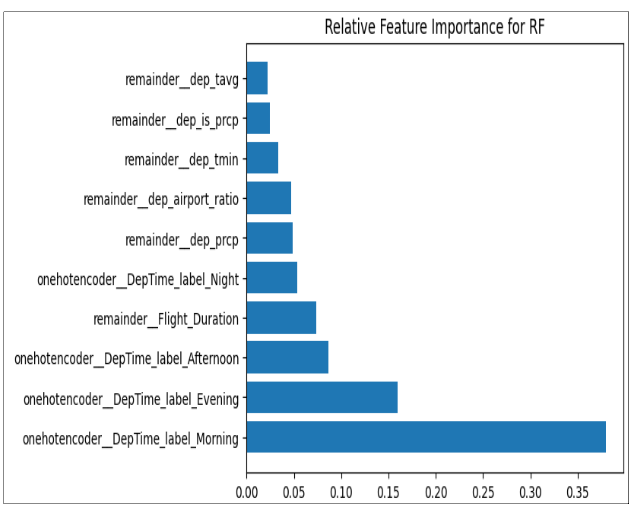


Figure 8. Top 10 Feature Importance for Random Forest

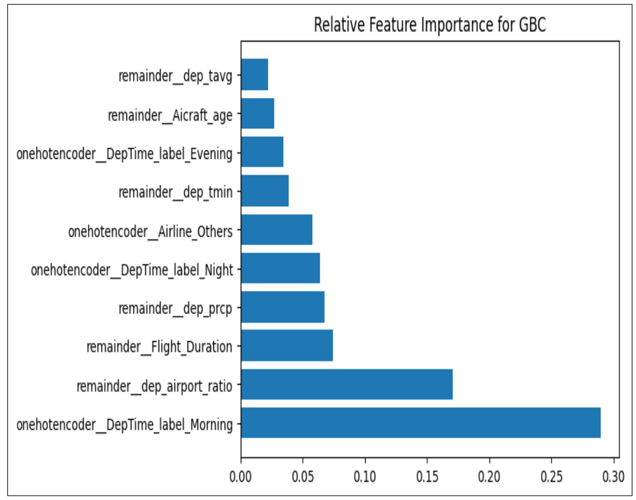


Figure 9. Top 10 Feature Importance for Gradient Boosting

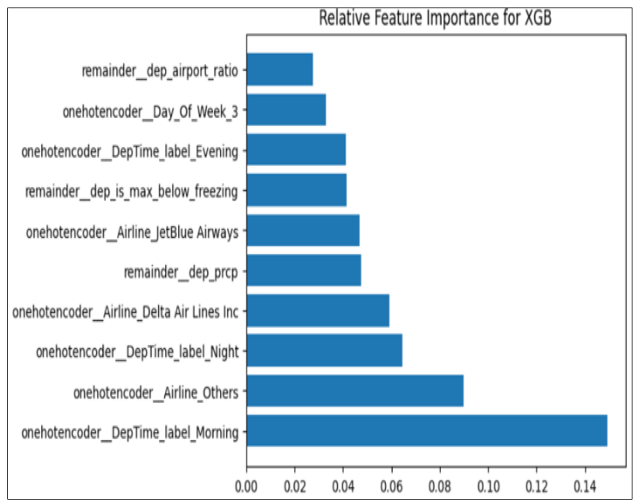


Figure 10. Top 10 Feature Importance for XGBoost

To further explain how the input features influence the model outputs, SHAP and LIME analyses are conducted on the best performing XGBoost model.

A global SHAP summary plot is generated with SHAP values, or Shapley Additive exPlanations breaking down the model's prediction to show the impact of each feature. The top features observed include precipitation, departure time, and temperature which are consistent with the findings in features importance analyses.

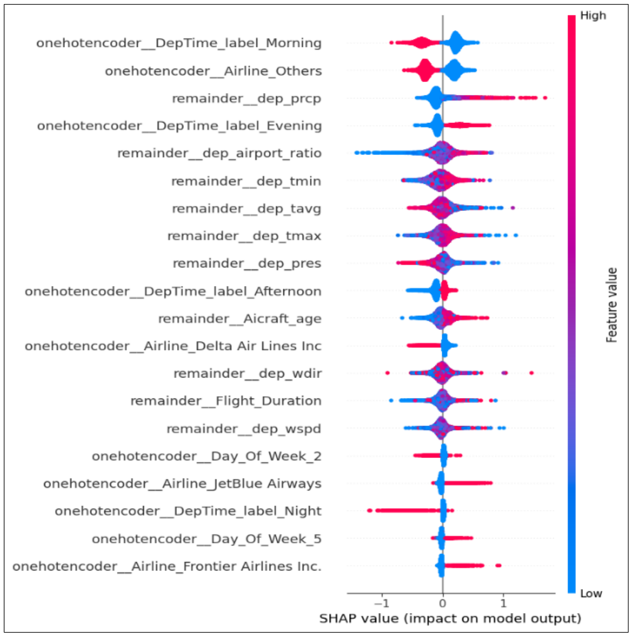


Figure 11. Result of feature influence in SHAP

Next, a localized SHAP force plot is used to examine how each feature influences the model prediction for a single flight instance. In the following example where a base log odd of severe delays at –2.63 lead up to a final prediction of –1.15 translates into a 24% probablity of a severe flightdelay derived using logistic function.

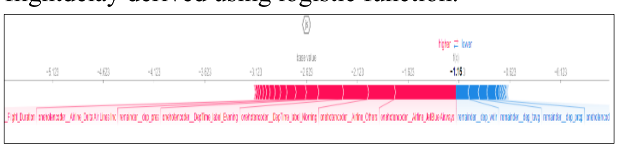


Figure 12. SHAP force plot of a single flight instance

Another way to explain the prediction probabilities of a single flight instance is by creating a LIME chart in which it seeks to create a simple model around the prediction that is easier to understand than the original model. In the following example, the probabiltity of xgboost model predicting the single flight instance being severely delay is 18% largely contributed by the departure time, airline type as well as the distance type.

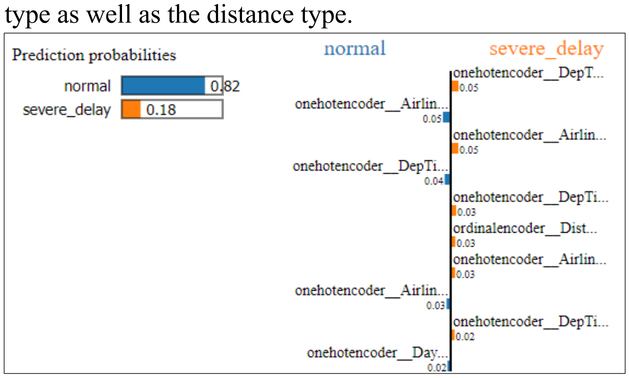


Figure 13. LIME chart a single flight instance

4. Translation to Business Impact

4.1 Explainable AI in Travel Insurance

This project has demonstrated the feasibility of ML to predict anomalies in flight schedules. For the travel insurance industry, the application of XAI is critical. Recognizing patterns in delays helps optimize customer service responses, such as allocating more resources during peak delay periods. This improves response times and customer satisfaction during critical periods. For example, with the information on bad weather conditions and takeoff timing, the company can communicate clearly to customers on the decisions made for the services and associated. Being alerted to potential disruptions will enhance trust and customer experience, differentiating the company from competitors and ultimately lead to higher customer retention.

4.2 Profitability as First Mover

A travel insurance company that adopts ML-driven insurance packages stands to benefit significantly by practicing price discrimination through dynamic pricing at individual flight level. It can better manage risk, reducing unexpected claim payouts and enhancing financial stability. Any deviation from industry pricing practice would likely be scrutinized by regulatory bodies. XAI helps ensure compliance by providing transparent and auditable explanations for decision-making in pricing. This first mover advantage also fosters innovation, allowing companies to set industry standards and capture market share. It is possible to replicate the success of predicting flights delays to the development of new insurance tailored to coverage for other types of disruptions such as cancellations or diversions.

5. Limitations and Deployment Considerations

5.1 Potential Limitations

While the data for this project coming from BTS may be adequate as proof of concept, to actualize the ML pipeline and generate ROI from its innovation phase, there need to understand the various stakeholders, from which their concerns and obstacles that need to be overcome.

To make data available for implementation, data owners and data custodians need to be distinguished. In the aviation industry, airlines would be the data owners of aircraft information while airports would take charge of the flights details, from schedules to disruptions. Significant effort is needed to communicate with both archetypes, so that everyone is aligned with the objectives and is willing to share information. A travel insurance company would likely act as the custodian of the data.

Data custodians must ensure confidentiality of information is enforced, balancing between explaining predictions of flight delays and privacy concerns. A particular aircraft can be undergoing quality issues, or that an airport is undergoing a major overhaul in its infrastructure. Such issues if kept in models can inadvertently propagate biases, leading to unfair or discriminatory outcomes and damaging reputations, even after the temporary issues are solved. The challenge in the long run is to differentiate what data is only useful in the short run.

5.2 Deployment

To effectively integrate the predictive model into an airline operation environment, it is crucial to establish a streamlined flow of information from upstream sources to downstream consumers. Furthermore, to maintain a competitive advantage, the frequency of model training is critical. The model should be updated regularly with new data to adapt to changing patterns in flight delays and operational conditions.

Upstream (Data Acquisition and Training)

For satisfactory model performance, there is a need to implement robust data cleaning and preprocessing techniques to handle missing values, outliers, and inconsistencies in the acquired data. Both the insurance company and data providers should have processes in place to ensure data quality and integrity, such as missing value handling techniques like imputation or interpolation. Also, leveraging domain knowledge about airlines, flights, and weather to guide data cleaning decisions would significantly aid to improve the quality and consistency of data.

It is also important to ensure compliance with data privacy regulations throughout the data acquisition, processing, and model training stages. Before acquiring data from airlines, ensure they have anonymized any passenger information that could be used to identify individuals. This might involve removing names, passport numbers, and other personally identifiable information. It is also crucial to train the model in secure computing environments that meet industry standards for data protection. This can further establish a data governance framework that outlines policies and procedures for data access, usage, and security throughout the model development lifecycle.

To ensure model fairness, it is crucial to identify potential biases in the training data that could skew the model's predictions. It is a common situation that models might overestimate delays for certain airlines if the training data is imbalanced; this can be mitigated using data augmentation techniques like oversampling, under sampling and synthetic data that address imbalances without directly copying existing data. Additionally, documenting efforts regarding identifying and mitigating bias in the model would boost transparency and build trust in the fairness of the model's predictions for insurance risk assessment.

Downstream (Model Deployment and Usage)

Firstly, it is imperative to continuously monitor the model's performance in production by establishing a feedback loop where the insurance company can provide insights on model performance in real-world scenarios. On top of tracking metrics curated towards delay predictions, data drift detection is another critical aspect to be considered; statistical tests or visualization techniques can assist in analyzing the logged data to identify if the distribution of the data used for prediction is shifting over time compared to the training data. This feedback can subsequently inform retraining decisions based on drift detection and identify significant performance degradation.

While it is useful to focus to monitor the model performance, the adaptability of the model to changing regulations within the industry would remain the most important downstream deployment consideration. Working with the insurance company to understand relevant regulations concerning model usage in the insurance sector would ensure that the model remains relevant to meet the business use case. This could require specific validation procedures and documentation requirements for the model before it can be used for insurance risk assessment effectively.

Model Retraining

To optimize the frequency of model retraining for predicting airline delays, several key factors must be considered:

- Data drift: Ever-evolving factors like airline policy shifts and volatile weather patterns can necessitate significantly more frequent updates for the model. Airline policies can undergo rapid revisions, and weather patterns can fluctuate wildly, even within short periods. These dynamic realities can quickly render a model's predictions outdated, potentially leading to costly inefficiencies or disruptions.
- Performance Monitoring: In the fast-paced world of airline industry, continuously monitoring prediction performance is paramount. Even slight degradations in accuracy, precision, or recall can have major downstream consequences for insurers. Inaccurate predictions can lead to airlines experiencing disruptions, cancellations, and delays, all of which translate into significant financial losses for insurance companies. Therefore, prompt retraining becomes essential to ensure the model's predictions remain reliable and insurers can make informed risk assessment decisions, potentially saving them millions.
- Cost and Resources: While retraining a model to address performance dips comes with substantial costs and resource allocation, the potential consequences of inaction for both airlines and insurers are far greater. Inaccurate predictions can lead to cascading disruptions for airlines, resulting in cancellations, delays, and reputational

damage. These translate into significant financial losses for insurers who underwrite those flights. Finding the right balance between retraining costs and maintaining a highly accurate model becomes crucial. By proactively addressing performance dips, insurers can potentially save millions through more accurate risk assessments and airlines can minimize disruptions, protecting their bottom line and passenger trust.

- Competitive Landscape: The competitive landscape within the airline insurance industry demands attention to how frequently competitors update their flight delay prediction models. Early detection of emerging trends or disruptions in the market can prompt a proactive retraining approach, allowing models and the respective insurance agents to stay ahead of the curve. This can benefit both airlines by minimizing delays and cancellations, and insurers by enabling them to offer competitive rates based on the most accurate risk assessments.

Effective strategies for determining retraining frequency include continuous data monitoring to detect distribution changes, performance tracking in real-time with alerts for falling metrics, and implementing a scheduled retraining framework, which could be quarterly or biannually. Moreover, trigger-based retraining could be employed to initiate updates when specific performance thresholds are crossed, or notable data drifts are detected. These approaches ensure the model remains robust and competitive in dynamically changing airline operation environments.

6. Conclusion

The objective of developing a predictive model to accurately forecast flight delays over 60 minutes aligns seamlessly with the overarching goal of enhancing airline operations and travel insurance services. By focusing on this specific delay threshold, the model directly addresses the most impactful travel disruptions, thereby providing significant value to both passengers and service providers. The strategic deployment of this model within airline operations, utilizing a continuous flow of data from upstream sources and delivering actionable insights to downstream consumers, displays its potential to significantly improve operational efficiency and customer satisfaction.

Maintaining the model's accuracy through regular retraining—guided by factors such as data drift, impact on prediction performance, cost considerations, and competitive dynamics—ensures that the model adapts to evolving conditions and maintains its relevancy and effectiveness. By employing strategies such as scheduled retraining and trigger-based updates, the model can respond dynamically to the changing environment, thus sustaining its competitive advantage.

In summary, this predictive model stands as a critical tool in the arsenal of airlines and travel insurers, poised to transform how delays are managed and enhance the travel experience by mitigating the negative impacts of flight disruptions.

References

US airline delay statistics - analyzed by TravelFreak. (2024, March 4). https://travelfreak.com/airline-delay-statistics/

Authors, A. Suppressed for anonymity, 2010.

Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

Kearns, M. J. *Computational Complexity of Machine Learning.* PhD thesis, Department of Computer Science, Harvard University, 1989.

Airline passengers worldwide by country. (2023, July 17). Statista Research Department. https://www.statista.com/statistics/537002/airline-passengers-worldwide-by-country/#statisticContainer

FlightDeckFriend.com. (2012-2022). Aircraft Maximum Wind Limits. Flight Deck Friend Limited. https://www.flightdeckfriend.com/ask-a-pilot/aircraft-maximum-wind-limits/

Dataset and Code for this Final Report. https://github.com/JaCh23/BT5153_AY2024_Group04