Advanced Modeling for Predicting Aviation Departure Delays: A Comprehensive Regression and Classification Approach

CSE 519 Data Science Fundamentals - Final Report

Aayush Nilesh Shah
Department of Computer Science
Stony Brook University
Stony Brook, NY
shah.aayush@stonybrook.edu

***Abstract*—This paper explores flight departure delays within the U.S. aviation system using a dataset spanning January 2018 to August 2023. The study integrates exploratory data analysis, geospatial insights, and machine learning. Key findings include correlations between airport busyness, airline size, airport terrain, and departure delays. Geospatial analysis identifies delay hotspots and assesses the impact of aviation entities on major airports. The methodology involves a novel neural network architecture and various models for regression and classification. Results showcase the importance of feature selection, with LightGBM Regression excelling in regression tasks and Random Forest in classification. The study contributes actionable insights for aviation management and efficiency.**

I. Introduction

The global aviation industry has seen tremendous growth in recent decades, enabling international travel, fostering economic connections, and connecting people globally. However, this expansion has brought a significant challenge: flight delays. Flight delays have far-reaching economic, environmental, and social consequences beyond inconveniencing travellers.

Economically, flight delays impose a heavy burden on airlines, passengers, and the broader economy. In 2022, U.S. passenger airlines spent an astonishing $101.18 per minute on aircraft block time, with fuel costs soaring by 87% to $42.15 per minute and labour costs rising by 3% to $28.99 per minute. Maintenance costs increased by 6%, resulting in billions of dollars in extra expenses for airlines. Additional costs are incurred in terms of extra gates, ground personnel, and operational resources. For passengers, flight delays cost them billions of dollars annually, assuming an average value of $47 per hour for their time [1]. In 2019, the Federal Aviation Administration estimated the annual cost of delays, including direct costs to airlines and passengers, lost demand, and indirect costs, at a staggering $33 billion [2].

Moreover, flight delays have a significant environmental impact, primarily in terms of increased CO2 emissions. In 2022, flight disruptions contributed approximately 3 million additional tons of CO2 emissions in the United States, 4-5 million tons in Europe, and less than 1 million tons in Australia. These cumulative emissions were equivalent to the annual emissions of approximately 2 million cars. Flight delays also lead to increased fuel consumption and generate up to 90,000 tons of waste annually due to unplanned hotel stays and meals. Additionally, they contribute to noise pollution, particularly in densely populated areas, which can lead to health issues for affected residents [3].

A. Motivation

Flight delays have far-reaching consequences, affecting the economy, the environment, and passenger well-being, underscoring the need for more sustainable aviation practices. Delays stem from various factors like air traffic control restrictions, adverse weather, and crew availability issues, making their prevention a crucial area of aviation research and innovation. Flight delays have broader consequences, impacting not only the economy and the environment but also passenger well-being and local communities. Travellers endure physical and mental strain from prolonged waiting times and disrupted plans, leading to increased stress. These delays also highlight the need for more sustainable aviation practices.

The causes of flight delays are diverse, including air traffic control restrictions, adverse weather, bird strikes, ripple effects from prior delays, strikes, connecting passenger wait times, and crew availability issues. With their far-reaching effects, addressing and preventing flight delays have become vital areas of research and innovation in the aviation industry.

B. Contributions

The notable contributions of this study are as follows:

- Investigated the impact of airport busyness on departure delays, revealing a direct relationship.
- Explored the correlation between airline size and median departure delays, highlighting operational efficiencies.
- Conducted geospatial analysis, pinpointing flight delay hotspots and regional variations.
- Utilized machine learning models for regression and classification, enhancing predictive capabilities.

- Developed a novel neural network architecture tailored for mixed data types, improving performance.
- Proposed a novel clustering-based categorical encoding algorithm for managing over 9000 unique categorical values, enhancing feature representation and controlling dimensional explosion.

II. Related Work

Extensive research on flight delay prediction has been driven by the increasing air traffic and the need for more accurate models to understand and mitigate delays. Notable studies include:

- Guo *et al.* [4], who introduced a hybrid model combining Random Forest Regression and the Maximal Information Coefficient, showing superior predictive accuracy compared to conventional methods, but lacking real-time weather data.
- Li *et al.* [5], who investigated weather-related and non-weather-related factors using a probabilistic Random Forest Model, enhancing accuracy but not considering dependencies between the categories.
- Anguita *et al.* [6], who performed a comparative analysis of various machine learning and deep learning models for flight departure time regression, revealing machine learning models' superiority but with data limitations.
- Mokhtarimousavi *et al.* [7], who employed a mixed logit model and Support Vector Machines with the Artificial Bee Colony algorithm to analyze flight delay factors using MIA data. The study focuses on methodology due to data constraints and suggests future research areas, such as seasonal variations and network-wide implementation.
- Bisandu *et al.* [8], who compared LSTM and BiLSTM for flight delay prediction, with BiLSTM outperforming LSTM. However, the dataset's imbalance may affect the accuracy metric's interpretation.

III. Dataset Description

This section provides an overview of the features used in our project through various data sources.

A. Flight Data Features

Our flight data originates from the United States Department of Transportation's Bureau of Transportation Statistics, covering the period from January 2018 to August 2023, comprising 100,584,764 flight records. These records represent extensive flight operations within the United States during this timeframe.

Our flight data contains various features, and descriptions for those features are provided in Table I.

B. Weather Features

The dataset[1] includes essential meteorological parameters, as listed in Table II.

[1]Sourced from National Oceanic and Atmospheric Administration (NOAA) and Germany's National Meteorological Service (DWD) through Meteostat API

TABLE I: Flight Data Features and Descriptions

Feature	Description
Year	Year of the flight
Month	Month of the flight
DayofMonth	Day of the month of the flight
DayOfWeek	Day of the week of the flight
FlightNumberMarketingAirline	Flight number by the marketing airline
TailNumber	Aircraft's tail number
Origin	Origin airport code
OriginCityName	Name of the origin city
Dest	Destination airport code
DestCityName	Name of the destination city
CRSDepTime	Scheduled departure time
DepTime	Actual departure time
TaxiOut	Time spent taxiing out before takeoff
TaxiIn	Time spent taxiing in after landing

TABLE II: Weather Features in the Dataset

Feature	Description
temp	Temperature
dwpt	Dew Point
rhum	Relative Humidity
prcp	Precipitation
wdir	Wind Direction
wspd	Wind Speed
pres	Sea-Level Air Pressure
coco	Weather Condition Code
snow	Snow Depth
wpgt	Wind Peak Gust
tsun	Total Sunshine Duration

C. Geospatial Data

The geospatial data[2] used is manipulated using the Geopandas library. The relevant columns in the dataset include `STATE_NAME` representing the full state name, `STATE_ABBR` for the state abbreviation, and `geometry` providing the geometric information for each state.

To investigate the impact of airspace congestion arising from the proximity of various aviation entities, additional data was procured[3]. This dataset encompasses pertinent information about aviation entities, including their type, name, elevation, continent, ISO country code, ISO region, municipality, GPS code, IATA code, local code, and coordinates. The inclusion of this data aims to explore the correlation between the spatial distribution of aviation facilities and departure delays, shedding light on potential congestion-related factors impacting flight punctuality.

IV. Exploratory Data Analysis

This section conducts the exploratory analysis, a pivotal step in constructing a robust flight data prediction model. We utilize informative graphs to offer useful insights. Throughout the analyses, the median departure delay serves as a central tendency measure, chosen for its resilience against outliers, enabling a more precise evaluation of typical departure delays.

[2]http://www.efrainmaps.es

[3]https://raw.githubusercontent.com/datasets/airport-codes/master/data/airport-codes.csv

A. *Effect of Airport Busyness on Departure Delays*

In this section, we analyze departure delays for US flights from January 2018 to August 2023, aiming to explore the connection between airport busyness and median departure delay.

Airports are classified into three categories based on their total flight count during this period. The top 25% are labelled ”Busiest,” the bottom 25% as ”Least Busy,” and the middle airports as ”Moderately Busy.”

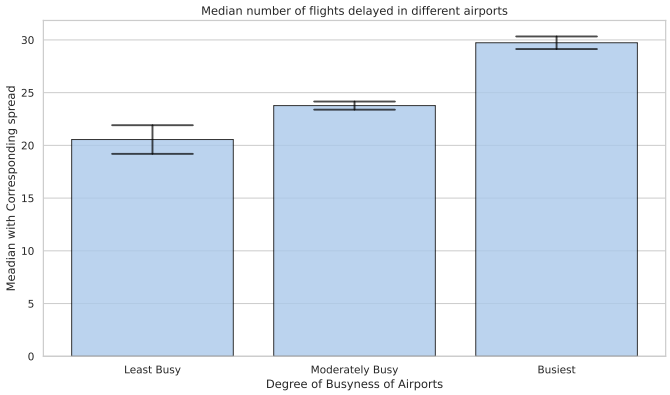


Fig. 1: Median Departure Delay vs. Degree of Busyness of Airports

Figure 1 depicts the connection between airport busyness and median departure delay. The x-axis categorizes busyness as ”Least Busy,” ”Moderately Busy,” and ”Busiest,” while the y-axis shows the median departure delay along with its spread. Our analysis reveals a clear trend: departure delay is directly proportional to airport busyness. As airports become busier, higher departure delays are observed. This trend is emphasized by the increasing spread of median departure delay values with the rising busyness of airports. This relationship is attributed to busier airports handling more flights and passengers, leading to increased congestion, longer taxi times, and a higher likelihood of delays. Additionally, more flights increase the potential for disruptions caused by factors like weather, air traffic control, and scheduling conflicts.

In conclusion, our analysis establishes a significant relationship between airport busyness and median departure delay. As airports become busier, departure delays tend to be longer, providing valuable insights for airline operations and passengers in managing expectations and travel plans.

B. *Effect of Airlines Size on Departure Delays*

In this section, we explore the relationship between airline size, as categorized by the 0.7 quantile threshold, and median departure delays for flights. The quantile method objectively determined the threshold, classifying airlines as ”large” or ”small” based on their relative counts in the dataset.

In Figure 2, the y-axis represents the median departure delay, while the x-axis categorizes airlines as ”Large” and ”Small” based on a 0.7 quantile threshold, identifying the top 30% with the highest counts. The plot reveals an inverse relationship between departure delay and airline size, with larger airlines exhibiting lower delays, indicating potential operational efficiencies. Smaller airlines, with fewer resources, show relatively higher median departure delays, suggesting operational challenges.

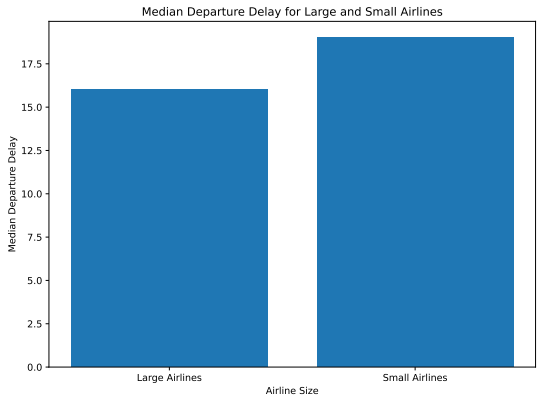


Fig. 2: Median Departure Delay vs. Airlines Size

In conclusion, this analysis establishes a correlation between airline size and median departure delays. Larger airlines, major players in air traffic, tend to have lower delays compared to their smaller counterparts, emphasizing the significance of strategies to improve flight punctuality and passenger satisfaction.

C. *Correlation of altitude with delays*

In this section, we explore the relationship between airport altitude and flight departure delays to better understand how an airport’s terrain might affect departure delays. The method employed involves clustering airports based on their altitudes using the K-Means algorithm. The Elbow Method is utilized to determine the optimal number of clusters (here 2), ensuring a meaningful categorization.

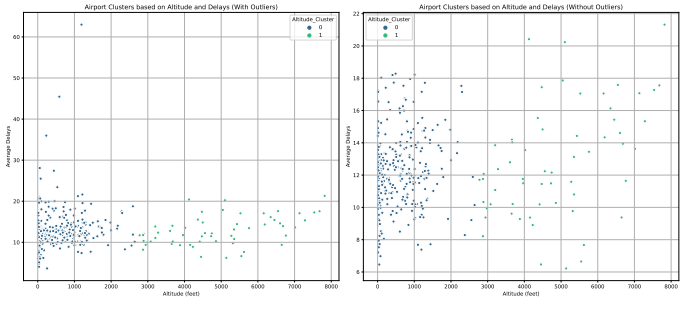


Fig. 3: Correlation of Altitude with Median Delays. i) With Outliers, ii) Without Outliers

To enhance the accuracy of our analysis, we address the impact of outliers, particularly in cluster one. Outliers are identified and removed to better understand the general trends within this cluster. This nuanced approach aims to mitigate the influence of extreme values on our observations, providing

a more accurate representation of the correlation between altitude and departure delays.

The results are visualized in Figure 3, with the first plot showcasing the clusters and their associated median delays, offering an initial overview of the dataset. Subsequently, a second plot is generated after removing outliers from Cluster 1, allowing for a more refined examination of altitude and departure delay correlations.

Within each cluster, the data shows a significant spread, indicating no clear correlation between altitude and delays at the same altitude. The wide dispersion suggests a lack of consistent patterns, and therefore we can conclude that altitude doesn't have a substantial effect on the departure delays.

V. Geospatial Data Analysis

Geospatial analysis is vital for studying flight delays, utilizing tools like heat maps to identify geographic patterns and regional trends. It supports targeted predictive modeling and allows customized strategies for high-risk locations. This approach enhances the understanding of delay distribution, providing insights to optimize operational efficiency, allocate resources strategically, and improve air travel system reliability. Figures exclude Alaska and Hawaii for clearer map visualization.

A. Hotspots of Flight Delays

In this section, we generate a heatmap to visually represent median flight delays across various geographic locations. The data includes all the flight data of 389 airports under consideration from January 2018 to August 2023.

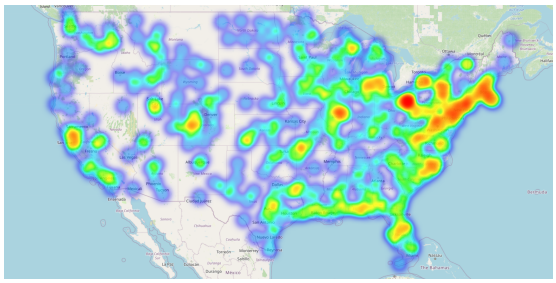


Fig. 4: Snapshot of the heatmap for median delays

Each color patch in Figure 4 corresponds to a group of airports in the highlighted area, with color intensity indicating the median delay. The gradient, from cooler to warmer tones, effectively communicates delay magnitude, with warmer shades representing longer delays.

Clusters of heightened color suggest regions with consistently longer median delays, pinpointing hotspots. Isolated areas of intense color indicate specific airports frequently experiencing substantial delays.

Specific regions in the US have more common departure delays, particularly concentrated on the East Coast. The East Coast, housing busy airports like in Washington DC and NYC, faces tougher weather conditions than the West Coast, contributing to flight delays.

B. Analysis of delays by state

In this section, we generate a heatmap for providing a visual representation of a comprehensive overview of the average median delays experienced across different states.

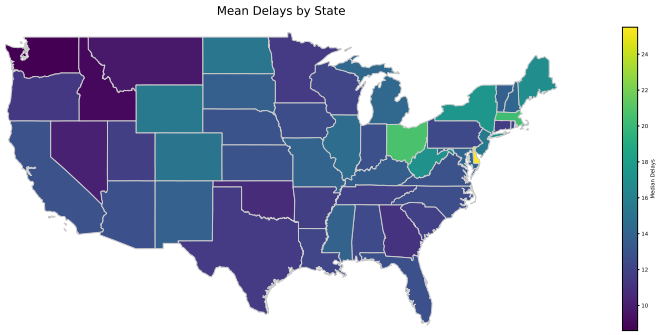


Fig. 5: Heatmap of Average Departure Delays by State

In Figure 5, color intensity corresponds to average delays, with warmer shades indicating higher delay values. The visualization links delays with specific states using spatial operations like 'contains,' summarizing cumulative delays for each state. Average delays are used instead of median to provide a more representative measure of central tendency, considering that median delays for each airport already account for outliers.

This plot assists in identifying states with consistently higher or lower average delays, offering insights into regional variations. Disparities in airport infrastructure, varying climates, and uneven air traffic distribution among states may contribute to these differences. Terrain disparities within states could also play a significant role in causing variations.

C. Proximity Analysis of Aviation Entities and Major Airports

In this section, we want to study the effect of congestion of airspace due to the proximity of aviation entities (Heliports, Seaplane Bases, Balloonports, and Small Airports) on the departure delays of flights departing from the major airports across the United States. To study the effect of these entities we have plotted a heatmap of median delays in each major airport along with aviation entities.

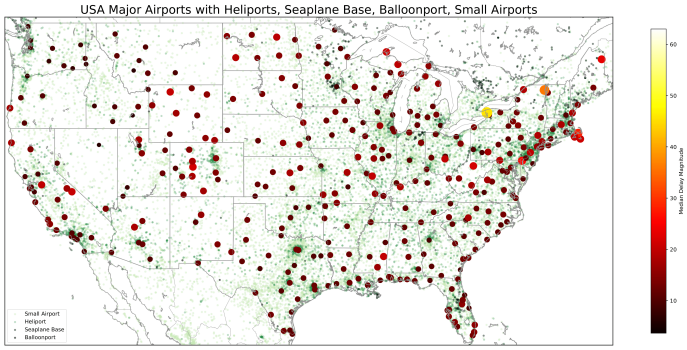


Fig. 6: Aviation space congestion visualization

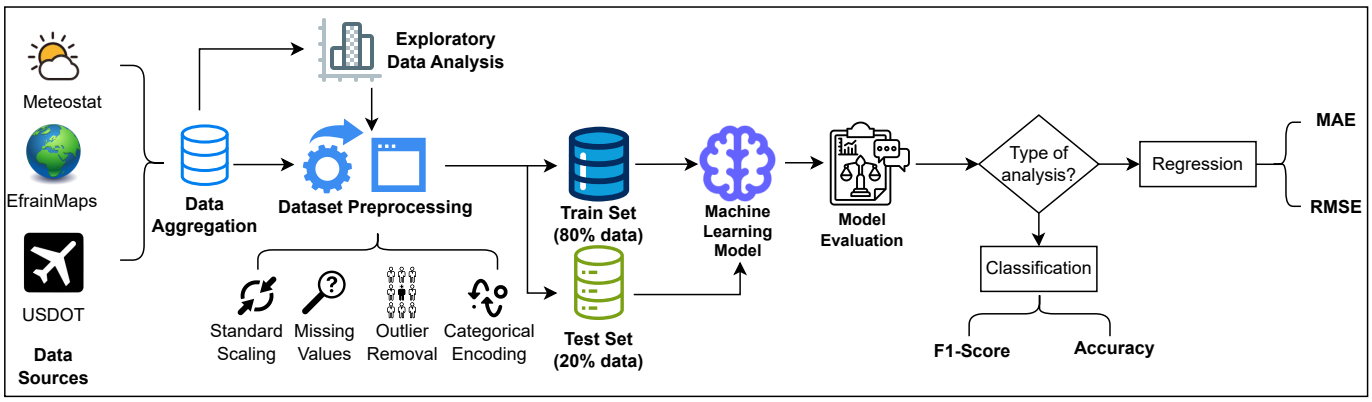


Fig. 7: System Model

To support the interpretation of this heatmap, we've quantified the number of aviation entities within a 45-mile radius of each major airport. This specific radius was chosen to account for shared airspace and potential interactions among various aviation facilities, providing a meaningful distance to assess their combined impact on departure delays at major airports. Spatial queries, such as 'buffer' and 'within,' have been employed to determine the count of entities.

The circles in Figure 6 symbolize major airports, with warmer and larger circles indicating higher median delays. The four distinct shades of green dots correspond to Heliports, Seaplane Bases, Balloonports, and Small Airports, respectively.

We can observe that there's a significant amount of congestion around almost all the major airports. Still, the amount of delay at each airport doesn't seem to follow the amount of congestion around it. We cannot conclude if congestion has a substantial effect on the departure delays of the major airports.

To support this observation, we computed the correlation between the number of entities surrounding an airport and the median delays associated with them. The resulting correlation value is 0.03, suggesting a positive relationship, but its significance is limited. Therefore, we can conclude that the presence of aviation facilities around major airports may not be a significant factor causing delays within these airports.

VI. Methodology

In this section, we describe the step-by-step process followed in our study, encompassing data preparation, feature selection, regression analysis, and classification analysis.

A. System Model

Figure 7 illustrates the sequential flow of our approach, encompassing diverse data sources, rigorous data preprocessing steps guided by insights from exploratory data analysis, a strategic train-test split, the subsequent training of machine learning models, and the final evaluation metrics.

For regression tasks, the evaluation includes Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) [9], while for classification tasks, performance is assessed using F1-Score and Accuracy metrics [10].

B. Data Preparation

1) Handling Missing Values: In addressing missing values within the dataset, it was observed that the columns snow, wpgt, and tsun exclusively contained null values. Consequently, these columns were deemed unsuitable for analysis and were consequently excluded from further consideration.

Regarding other columns, specifically temp and dwpt, where the occurrence of null values surpassed 300k instances, an interpolation strategy was attempted. The methodology involved utilizing the rolling mean of either preceding or succeeding values to estimate the missing data points. Despite employing an extensive window for the rolling mean, the attempted interpolation proved ineffective, revealing a pronounced clustering of missing values. Given the inherent limitations of interpolation in such clustered scenarios, the decision was made to exclude these rows from the dataset.

Additionally, for columns with a relatively lower count of null values, such as prcp, null values were substituted with 0, indicating an absence of precipitation during the corresponding time intervals. This treatment ensured the retention of valuable data while mitigating the impact of missing values on subsequent analyses.

2) Standardization: In addressing the diverse scales present in numerical features, it became crucial to implement standardization for meaningful comparisons. Numerical features exhibited significant variations in magnitudes, prompting the application of the z-score transformation [11]. This involved subtracting the mean (μ) and dividing by the standard deviation (σ), resulting in a standardized distribution with a mean of zero and a standard deviation of one:

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

where Z is the standardized value, X is the original value, μ is the mean of the feature, and σ is the standard deviation of the feature.

By standardizing the dataset, we eliminate the scale-related disparities, enabling fair and unbiased comparisons among features. This ensures that each feature contributes proportionally to the analysis, regardless of its original scale, thereby

enhancing the interpretability and performance of subsequent modeling and analytical procedures.

3) Categorical Encoding: To address the challenge of handling over 9000 unique categorical values within our dataset, we devised a distinctive strategy outlined in the Algorithm 1 for managing categorical features.

Algorithm 1: Categorical Clustering Algorithm

Data: Input DataFrame D, Categorical Columns C, Target Variable T
Result: Clustered DataFrame C_D

foreach *cat_column in C* **do**
  $unique\_values \leftarrow D[cat\_column].unique()$;
  $agg\_df \leftarrow$ CreateEmptyDataFrame($unique\_values$, ['Agg_DepDelay']);
  foreach *value in unique_values* **do**
    $agg\_df.loc[value,'Agg\_DepDelay'] \leftarrow$ AggregateMedian($D[D[cat\_column] == value][T]$);
  $agg\_df['Agg\_DepDelay'] \leftarrow$ Standardize($agg\_df[['Agg\_DepDelay']]$);
  $distortions \leftarrow []$;
  $K\_range \leftarrow$ [1, 2, ..., 10];
  foreach k *in* K_range **do**
    $kmeans \leftarrow$ FitKMeans($agg\_df[['Agg\_DepDelay']], k$);
    $distortions.append(kmeans.inertia)$;
  $kneedle \leftarrow$ FindElbowPoint($K\_range, distortions$);
  $optimal\_k \leftarrow kneedle.elbow$;
  $kmeans \leftarrow$ FitKMeans($agg\_df[['Agg\_DepDelay']], optimal\_k+1$);
  $agg\_df['Cluster'] \leftarrow$ PredictClusters($kmeans, agg\_df[['Agg\_DepDelay']]$);
  $C_D[cat\_column] \leftarrow C_D[cat\_column].map(agg\_df['Cluster'].to\_dict())$;
$C_D \leftarrow$ OneHotEncode(C_D, C);

4) Outlier Removal: Outliers, defined as values exceeding 3 standard deviations from the mean, were identified and subsequently removed to enhance model performance as explained in Algorithm 2 [12]. This process is justified as outliers can significantly skew statistical analyses and modeling outcomes, leading to inaccurate predictions and reduced model reliability.

Eliminating outliers helps ensure the robustness and validity of the model by mitigating the influence of extreme values that could otherwise distort the overall pattern and trends within the dataset.

Following the comprehensive treatment of missing values and the removal of outliers, the dataset was refined to comprise a total of 85,724,944 rows.

Algorithm 2: Outlier Removal Algorithm

1: Calculate the mean (μ) and standard deviation (σ) for each feature in D.
2: Define the outlier threshold: $T = 3 \cdot \sigma$.
3: Identify outliers: $O = \{x \mid x \in D, x > \mu + T \text{ or } x < \mu - T\}$.
4: Remove outliers from D: $D' = D \setminus O$.
5: **return** D'

C. Feature Selection

Recognizing a significant correlation between two pairs of features, we strategically removed one feature from each correlated pair to mitigate multicollinearity. This reduction in correlated features addresses instability and inflated coefficients, ultimately enhancing the model's interpretability and generalization. This step ensures that the model remains unbiased and reliable by avoiding undue influence from redundant information, contributing to more dependable predictions. The correlation patterns are depicted in Figure 8 before and after removing correlated columns.

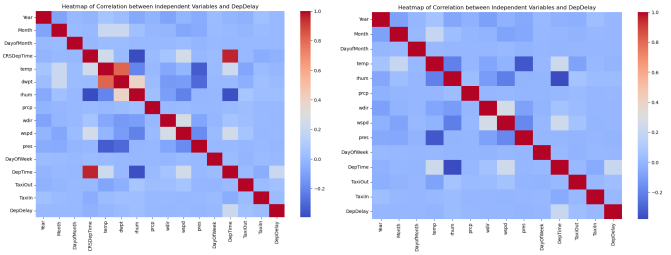


Fig. 8: Correlation Heatmap Before and After Removing Correlated Features

To identify the important features shaping our analysis and model development, we trained a baseline linear regression model. The model leveraged regression coefficients to quantify feature importance, as illustrated in Figure 9, providing a visual representation of each feature's relative significance. This preliminary analysis laid the groundwork for identifying and prioritizing influential features. We used the ten most important features for subsequent phases of our research.

D. Regression Analysis

While conducting regression analysis, we have used a novel neural network architecture that is designed for mixed data types, featuring multiple input layers for categorical features and numerical features. The embedding layers are instrumental in converting categorical variables into dense vectors, facilitating the model's understanding of relationships and patterns. The outputs from these embedding layers, along with numerical features, are concatenated and processed through several dense layers to produce a single output neuron. The embedding layers play a pivotal role in dimensionality reduction, capturing semantic relationships and enhancing the

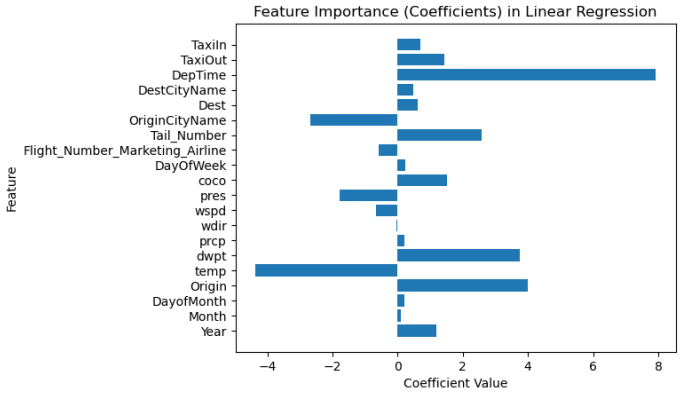


Fig. 9: Feature Importance

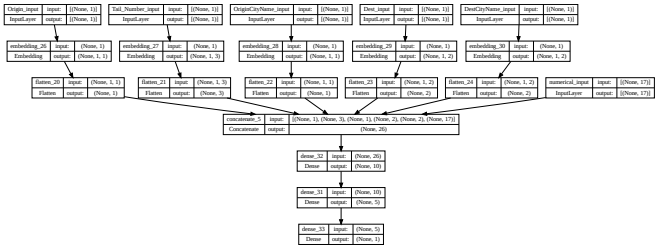


Fig. 10: Neural Network Architecture

model's ability to learn from both categorical and numerical data. The architecture of the neural network used is visualized in Figure 10.

We employed various other regression models to predict flight departure delay, including Linear Regression, Lasso Regression, Ridge Regression, ElasticNet Regression, and LightGBM Regression. The performance metrics for each model, both with and without outlier removal (OR), are presented in Table III.

TABLE III: Regression Model Performance

Model	Without OR		With OR	
	MAE	RMSE	MAE	RMSE
Linear Regression	16.64	34.98	14.50	35.37
Lasso Regression	16.65	35.13	14.75	35.63
Ridge Regression	16.64	34.98	14.50	35.37
ElasticNet Regression	16.79	35.19	14.80	35.66
Neural Network	16.34	34.97	14.21	34.23
LightGBM Regression	15.50	34.62	**13.83**	**33.64**

The LightGBM Regressor outperformed other regression models due to its ability to handle complex relationships and non-linearities in the dataset. Flight departure delay prediction involves intricate patterns influenced by various factors such as weather conditions, airport congestion, and air traffic control, which LightGBM's tree-based approach can effectively capture.

The improvement in MAE after outlier removal can be attributed to the removal of extreme values that might have disproportionately influenced the mean absolute error. Outliers, such as highly delayed flights or data recording errors, can distort the evaluation metric, and their removal ensures a more representative evaluation of model performance.

To further enhance model performance, we employed Recursive Feature Elimination (RFE) to refine the feature subsets for various regression algorithms. The number of features to be selected in RFE was set to five. The results obtained for each model after employing RFE are shown in Table IV.

TABLE IV: Regression Model Performance with Outlier Removal and RFE

Model	MAE	RMSE
Linear Regression	13.94	34.26
Lasso Regression	14.41	35.34
Ridge Regression	13.94	34.26
ElasticNet Regression	14.57	35.42
LightGBM Regression	**12.36**	**32.95**

The outcomes reveal notable improvements in model performance across the board. Linear Regression, Lasso Regression, and Ridge Regression demonstrate reductions in both MAE and RMSE, indicating enhanced predictive power. ElasticNet Regression also benefits from RFE, showcasing a decrease in MAE and RMSE.

However, the most significant performance boost is observed in the LightGBM Regression model. With a lower MAE of 12.36 and RMSE of 32.95, the LightGBM model stands out as the top-performing regression model after the implementation of RFE. This suggests that the RFE process successfully identified and retained the most relevant features, contributing to the model's superior predictive performance.

These results underscore the importance of feature selection techniques such as RFE in refining models and improving their predictive power. The optimized feature subsets allow the models to focus on the most relevant information, resulting in more effective predictions and better overall performance.

E. Classification Analysis

To make the dataset suitable for classification, the target variable i.e. DepDelay was transformed to binary values as represented by the following mathematical equation:

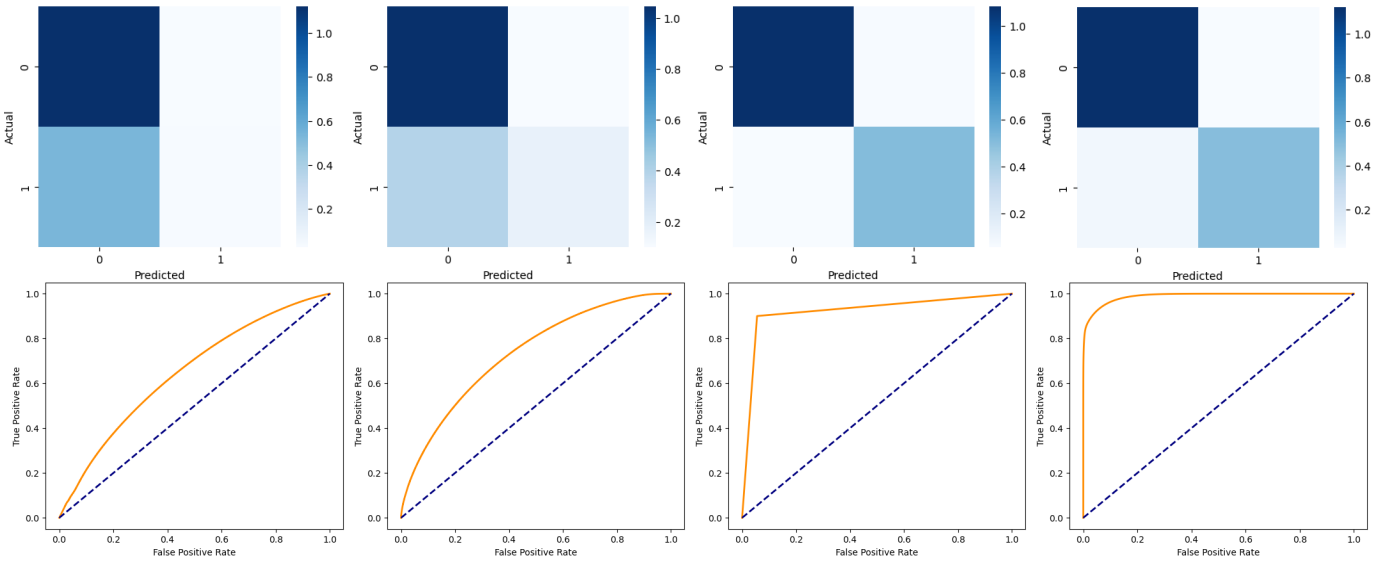
$$\text{DepDelay\_Binary} = \begin{cases} 0, & \text{if DepDelay} \leq 0 \\ 1, & \text{if DepDelay} > 0 \end{cases}$$

This transformation facilitated the classification task by categorizing flights as either delayed (1) or not delayed (0).

We applied various classification algorithms, including Gaussian Naive Bayes, LightGBM Classifier, Decision Tree, and Random Forest, to predict flight delays. The classification was performed on a dataset that had outliers removed and utilized clustered categorical encoding.

The results for each classification algorithm are presented in Table V, and class metrics, including confusion matrices and ROC curves, are depicted in Figure 11.

Random Forest outperformed other classifiers, achieving the highest F1 Score and Accuracy. The superiority of Random



(a) Gaussian Naïve Bayes (b) LightGBM Classifier (c) Decision Tree (d) Random Forest Classifier

Fig. 11: Confusion matrices and ROC curves of various machine learning algorithms: Gaussian Naïve Bayes, LightGBM Classifier, Decision Tree, and Random Forest Classifier.

TABLE V: Classification Results

Algorithm	F1 Score	Accuracy
Gaussian Naive Bayes	0.10	0.67
LightGBM Classifier	0.41	0.71
Decision Tree	0.91	0.87
Random Forest Classifier	**0.96**	**0.95**

Forest can be attributed to its ensemble nature, which combines multiple decision trees, capturing complex relationships in the dataset. This is especially advantageous in predicting flight delays, where numerous factors contribute to the outcome.

The confusion matrices and ROC curves in Figure 11 further highlight the superior performance of Random Forest. The higher Area Under the ROC Curve (AUC) indicates better discrimination between classes, reinforcing the robustness of Random Forest in distinguishing between delayed and non-delayed flights.

VII. Future Work

Our research has opened avenues for future exploration and improvement in flight departure delay prediction.

- Further refinement of the novel neural network architecture by exploring advanced techniques, such as attention mechanisms, to capture more intricate patterns in the data.
- Optimization and extension of the clustering-based categorical encoding algorithm to accommodate evolving datasets.
- Deeper exploration into the impact of specific weather conditions on departure delays, incorporating real-time data for more accurate predictions.
- Integration of advanced machine learning models and exploration of explainable AI techniques to enhance interpretability and facilitate broader industry adoption.
- Collaborative efforts with aviation stakeholders to implement and validate predictive models in real-world scenarios, contributing to the practical applicability of our findings.

VIII. Conclusion

In conclusion, our research illuminates key factors influencing flight departure delays in the United States. Our proposed clustering-based categorical encoding algorithm provides a robust solution for handling diverse categorical data. Insights into the correlation between airport busyness, airline size, airport terrain, and departure delays offer valuable considerations for operational improvements. Geospatial analysis identifies hotspots of delays, enabling targeted strategies for resource allocation. This study can be used further to optimize the economics of airlines and improve passenger experiences.

References

[1] *U.S. Passenger Carrier Delay Costs - Airlines For America*, May 2023.
[2] *Cost of delay estimates 2019 - Federal Aviation Administration*, Jul 2020.
[3] *Cost of disrupted flights to the economy - AirHelp*, Sep 2023.
[4] Z. Guo, B. Yu, M. Hao, W. Wang, Y. Jiang, and F. Zong, "A novel hybrid method for flight departure delay prediction using random forest regression and maximal information coefficient," *Aerospace Science and Technology*, vol. 116, p. 106822, 2021.
[5] Q. Li, R. Jing, and Z. S. Dong, "Flight delay prediction with priority information of weather and non-weather features," *IEEE Transactions on Intelligent Transportation Systems*, 2023.
[6] J. G. M. Anguita and O. D. Olariaga, "Flight departure delay forecasting," *Journal of Airport Management*, vol. 17, no. 2, pp. 197–209, 2023.
[7] S. Mokhtarimousavi and A. Mehrabi, "Flight delay causality: Machine learning technique in conjunction with random parameter statistical analysis," *International Journal of Transportation Science and Technology*, vol. 12, no. 1, pp. 230–244, 2023.

[8] D. B. Bisandu PhD and I. Moulitsas PhD, "A deep bilstm machine learning method for flight delay prediction classification," *Journal of Aviation/Aerospace Education & Research*, vol. 32, no. 2, p. 4, 2023.

[9] A. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," *arXiv preprint arXiv:1809.03006*, 2018.

[10] Ž. Vujović *et al.*, "Classification model evaluation metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021.

[11] N. Fei, Y. Gao, Z. Lu, and T. Xiang, "Z-score normalization, hubness, and few-shot learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 142–151, 2021.

[12] V. Barnett, T. Lewis, *et al.*, *Outliers in statistical data*, vol. 3. Wiley New York, 1994.