Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/55/20241522

Research on key factors of flight delays in the United States based on data mining

Zicheng Wang

Tiangong Innovation School, Tiangong University, Tianjin, 300387, China

stra_wzc@126.com

Abstract. Flight delays in the United States pose a significant global challenge. With the continuous growth of the aviation industry, the increasing number of flights raises demands on transportation infrastructure, making flight delay a serious challenge affecting the aviation industry and passengers. Through comparative analysis, we found that the average flight duration and departure times of delayed flights were significantly later than those of non-delayed flights. Additionally, the delay rates were highest in California and lowest in Texas for both departure and arrival locations. Using cluster analysis, major airlines in the United States were classified into three categories. Factor analysis was employed to analyse the correlations among different factors. Logistic regression revealed a positive correlation between departure times, flight durations, and flight delays. Conversely, the day of the week showed a negative correlation with flight delays. These studies provide practical insights for improving flight punctuality and enhancing the aviation transportation system. This, in turn, aids airlines in optimizing operations and mitigating the adverse impacts of delays on the economy and passengers.

Keywords: comparative analysis, cluster analysis, factor analysis, logistic regression.

1. Introduction

The aviation industry has long been regarded as one of the most vital sectors in modern society, connecting diverse regions and cultures while providing robust support for global economic development. Over the years, the aviation industry has experienced steady growth, averaging approximately 5% per annum over the past three decades [1]. As shown in Figure 1, except for the year 2020, which was significantly impacted by the COVID-19 pandemic, the United States has witnessed a continuous increase in air passenger traffic year after year [2]. As one of the world's largest aviation markets, the United States faces a prominent issue of flight delays. This problem encompasses not only technical challenges but also extends its ramifications to various economic and social domains. Flight delays have far-reaching consequences, affecting multiple sectors. For passengers, they disrupt original plans and schedules, resulting in additional financial burdens and psychological stress [3-5]. Airlines, on the other hand, bear the economic losses and reputation damage associated with delays, leading to increased operational costs. Economically, delays have repercussions on tourism, business activities, and freight transportation, among other aspects [6,7]. Societally, delays impact the environment, employment, and more.

DOI: 10.54254/2755-2721/55/20241522

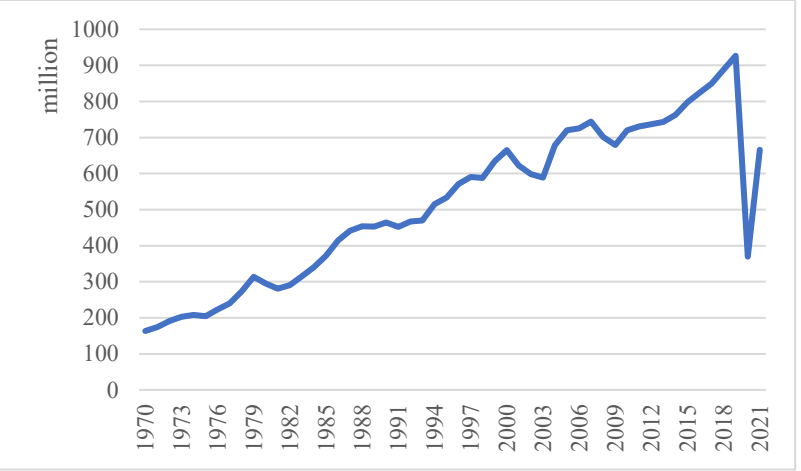

Figure 1. Air passenger volume in the United States over the years.

The issue of flight delays represents a global challenge [8,9], impacting not only the United States but also involving aviation transportation systems across various countries and regions worldwide. As the demand for global travel continues to rise, the increasing volume of flights exerts pressure on transportation infrastructures, consequently elevating the risk of delays. Given the multifaceted nature of this problem, a thorough investigation into the interplay of various factors and their relationship to flight delays is imperative. Such a comprehensive analysis seeks to yield valuable insights, improving punctuality, reducing economic costs, and fostering sustainability within the global aviation industry. In-depth research is pivotal in identifying more effective solutions to better serve the needs of both travelers and economic systems.

Many attempts have been by researchers in the past for predicting flight delays. Kim et al. [10] implemented a deep learning approach using recurrent neural networks (RNNs) to forecast flight delays. Ding et al. [11] presented a method for simulating arrival flights and a multilinear regression algorithm to forecast delays. Nigam et al. [12] employed logistic regression to combine weather data with airport information for predicting departure time delays. Manna et al. [13] established an accurate prediction model for both arrival and departure delays of flights by applying gradient-boosted decision trees. Chakrabarty et al. [14] proposed a machine learning model using a gradient boosting classifier to predict arrival delays of American airline flights at the five busiest airports in the United States.

In this paper, unlike conventional studies that primarily focus on forecasting flight delays, the emphasis is placed on an investigation of the primary factors influencing flight delays and the relationships among these factors. The goal is to facilitate future improvements in the aviation system and the reduction of delay rates. The methodology employed involves a comparative analysis, revealing significant associations between various flight characteristics and whether delays occur. Key findings include a disproportionate representation of WN flights in the delayed group, while DL flights dominate the non-delayed group. Additionally, there are noteworthy correlations between departure dates and the occurrence of delays, with Wednesdays having the highest proportion of delays among delayed flights and Thursdays among non-delayed flights. Furthermore, significant disparities are observed in average flight durations and departure times between the delayed and non-delayed groups, with delayed flights experiencing both longer average flight times and later departure times. Moreover, the location of the origin and destination states exhibits a significant relationship with the occurrence of delays, with California showing the highest delay rate and Texas the lowest. Employing cluster analysis, major U.S. airlines are categorized into three groups based on departure times and flight durations. Factor analysis is employed to reduce dimensionality in continuous data, analyze their interrelatedness. Logistic regression is employed to examine the relationships between delay occurrence and various factors, revealing positive associations with departure times and flight durations, as well as negative associations with departure times during specific time intervals within a week.

DOI: 10.54254/2755-2721/55/20241522

The rest of this paper is organized as follows. Section 2 introduces the basic information of the dataset. In Section 3, we use comparative analysis to analyze the significant relationship between various factors and delays. In Section 4, We categorized American airlines using cluster analysis. In Section 5, factor analysis was employed to explore the correlation among variables. In section 6, We used logistic regression analysis to examine the correlation between various factors and flight delays.

2. Dataset Introduction

The dataset is sourced from Kaggle[15]. As shown in Table 1, the dataset comprises information such as airline, flight number, departure station, destination station, departure date (1-7 representing the day of the week), departure time (measured in minutes from midnight), flight route length, and whether the flight was delayed (with two values, 0 for no delay and 1 for delay). Since there are no missing values in the dataset, it contains 539,383 samples with a total of 8 attributes. Among these attributes, Time and Length are continuous, while the remaining six are categorical.

Table 1. Feature Study

ID	Attribute/Feature Name	Attribute Type
F1	Airline	Categorical
F2	Flight	Categorical
F3	AirportFrom	Categorical
F4	AirportTo	Categorical
F5	DayOfWeek	Categorical
F6	Time	Continuous
F7	Length	Continuous
F8	Delay	Categorical

3. Comparative analysis

In this section, we conducted a detailed analysis of the significance of various factors in relation to flight delays.

3.1. The Significant Relationship Between Different Flights and Flight Delays

Following the construction of a contingency table and subsequent chi-square test, Table 2 and Table 3 shows a highly significant two-tailed p-value of 0.000, underscoring a significant discrepancy between flights categorized as delayed and those categorized as non-delayed. Within the delayed flights, Southwest Airlines (WN) had the highest proportion, constituting 27.3%. Conversely, among the non-delayed flights, Delta Air Lines (DL) held the majority share at 11.2%. Importantly, a more in-depth analysis revealed a notably higher rate of delays for flights operated by airlines based in the western region compared to those based in the eastern region.

Table 2. Delay * Airline Crosstabulation

			9E	…	DL	…	WN	…	YV	Total
Delay	0	Count	12460	…	33488	…	28440	…	10391	299119
		% within Delay	4.2%	…	11.2%	…	9.5%	…	3.5%	100.0%
		% within Airline	60.2%	…	55.0%	…	30.2%	…	75.7%	55.5%
	1	Count	8226	…	27452	…	65657	…	3334	240264
		% within Delay	3.4%	…	11.4%	…	27.3%	…	1.4%	100.0%
		% within Airline	39.8%	…	45.0%	…	69.8%	…	24.3%	44.5%

DOI: 10.54254/2755-2721/55/20241522

Table 2. (continued).

Total	Count	20686	…	60940	…	94097	…	13725	539383
	% within Delay	3.8%	…	11.3%	…	17.4%	…	2.5%	100.0%
	% within Airline	100.0%	…	100.0%	…	100.0%	…	100.0%	100.0%

Table 3. Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	38193.571[a]	17	0.000
Likelihood Ratio	38787.957	17	0.000
N of Valid Cases	539383		

3.2. The Significant Relationship Between Departure Date and Flight Delays

Upon constructing a contingency table and conducting a chi-square test, Table 4 and Table 5 obtained a remarkably significant p-value of 0.000, indicating a substantial disparity in departure dates between the delayed and non-delayed categories. Notably, Wednesday departures constituted the majority among delayed flights, comprising 17.6% of the total. Conversely, Thursday departures were most prevalent among non-delayed flights, representing a share of 16.8%.

Table 4. Delay * DayOfWeek Crosstabulation

			1	…	3	4	…	7	Total
Delay	0	Count	38739	…	47492	50201	…	38186	299119
		% within Delay	13.0%	…	15.9%	16.8%	…	12.8%	100.0%
		% within DayOfWeek	53.2%	…	52.9%	54.9%	…	54.6%	55.5%
	1	Count	34030	…	42254	41244	…	31693	240264
		% within Delay	14.2%	…	17.6%	17.2%	…	13.2%	100.0%
		% within DayOfWeek	46.8%	…	47.1%	45.1%	…	45.4%	44.5%
Total		Count	72769	…	89746	91445	…	69879	539383
		% within Delay	13.5%	…	16.6%	17.0%	…	13.0%	100.0%
		% within DayOfWeek	100.0%	…	100.0%	100.0%	…	100.0%	100.0%

Table 5. Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1178.121[a]	6	0.000
Likelihood Ratio	1182.169	6	0.000
Linear-by-Linear Association	370.233	1	0.000
N of Valid Cases	539383		

3.3. The Significant Relationship Between Flight Duration and Flight Delays

In Table 6, we can see that the average flight duration for non-delayed flights was 129.66 minutes, whereas for delayed flights, it was 135.37 minutes. We observed that the average flight duration for

DOI: 10.54254/2755-2721/55/20241522

delayed flights was significantly longer than that for non-delayed flights. In Table 7, normality tests were conducted on the 'Length' data. The p-values for both the delayed and non-delayed groups were less than 0.05, rejecting the assumption of normal distribution. Thus, non-parametric tests were employed. The null hypothesis stated that there was no significant difference in flight duration ('Length') between delayed and non-delayed flight groups. Table 8 shows that according to the independent samples Mann-Whitney U test, the significance value was 0.00, clearly rejecting the null hypothesis. This rejection indicates a significant difference in flight duration ('Length') between the delayed and non-delayed flight groups.

Table 6. Descriptives

Delay				Statistic	Std. Error
Length	0	Mean		129.66	0.126
		95% Confidence Interval for Mean	Lower Bound	129.41	
			Upper Bound	129.90	
	1	Mean		135.37	0.146
		95% Confidence Interval for Mean	Lower Bound	135.08	
			Upper Bound	135.66	

Table 7. Tests of Normality

Delay		Kolmogorov-Smirnov[a]		
		Statistic	df	Sig.
Length	0	0.115	299119	0.000
	1	0.115	240264	0.000

Table 8. Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Length is the same across categories of Delay.	Independent-Samples Mann-Whitney U Test	0.000	Reject the null hypothesis.

3.4. The Significant Relationship Between Departure Time and Flight Delays

In Table 9, the average departure time for non-delayed flights was 12:45, whereas for delayed flights, it was 14:09. We observed that the average departure time for non-delayed flights was significantly earlier than that for delayed flights. In Table 10, normality tests were conducted on the 'time' data. The p-values for both the delayed and non-delayed groups were less than 0.05, rejecting the assumption of normal distribution. Non-parametric tests were therefore employed. The null hypothesis stated that there was no significant difference in departure time ('Time') between delayed and non-delayed flight groups. Table 11 shows that according to the independent samples Mann-Whitney U test, the significance value was 0.00, clearly rejecting the null hypothesis. This rejection indicates a significant difference in departure time ('Time') between the delayed and non-delayed flight groups.

DOI: 10.54254/2755-2721/55/20241522

Table 9. Descriptives

Delay				Statistic	Std. Error
Time	0	Mean		765.24	0.519
		95% Confidence Interval for Mean	Lower Bound	764.22	
			Upper Bound	766.25	
	1	Mean		849.41	0.538
		95% Confidence Interval for Mean	Lower Bound	848.35	
			Upper Bound	850.46	

Table 10. Tests of Normality

Delay		Kolmogorov-Smirnov[a]		
		Statistic	df	Sig.
Length	0	0.115	299119	0.000
	1	0.115	240264	0.000

Table 11. Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of Time is the same across categories of Delay.	Independent-Samples Mann-Whitney U Test	0.000	Reject the null hypothesis.

3.5. The Significant Relationship between Departure States and Flight Delays

The numbers 1-51 correspond to the alphabetical order of the states in the United States, with 52 representing non-contiguous states.

As shown in Table 12 and Table 13, a cross-tabulation and chi-squared test revealed a two-tailed asymptotic significance level of 0.000, indicating a significant difference between the delayed and non-delayed flights based on the departure state. Among delayed flights, the highest proportion of departures were from California, accounting for 13.0%. In contrast, among non-delayed flights, the majority of departures were from Texas, constituting 10.7%.

Table 12. Delay * StateFrom Crosstabulation

			1.00	…	5.00	…	43.00	…	52.00	Total
Delay	0	Count	2354	…	28214	…	32000	…	1442	299119
	1	Count	1319	…	31167	…	27300	…	952	240264
Total		Count	3673	…	59381	…	59300	…	2394	539383

Table 13. Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	8700.620[a]	50	0.000
Likelihood Ratio	8768.074	50	0.000
Linear-by-Linear Association	236.421	1	0.000
N of Valid Cases	539383		

DOI: 10.54254/2755-2721/55/20241522

3.6. The Significant Relationship between Arrival States and Flight Delays

In accordance with the data presented in Table 14 and Table 15, a cross-tabulation and chi-squared test revealed a two-tailed asymptotic significance level of 0.000, indicating a significant difference between the delayed and non-delayed flights based on the arrival state. Among delayed flights, the highest proportion of arrivals were in California, accounting for 13.1%. In contrast, among non-delayed flights, the majority of arrivals were in Texas, constituting 11.4%.

Table 14. Delay * StateTo Crosstabulation

			1.00	…	5.00	…	43.00	…	52.00	Total
Delay	0	Count	2119	…	27798	…	34014	…	1143	299119
	1	Count	1603	…	31553	…	25276	…	1260	240264
Total		Count	3722	…	59351	…	59290	…	2403	539383

Table 15. Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	7312.299[a]	50	0.000
Likelihood Ratio	7341.980	50	0.000
Linear-by-Linear Association	443.173	1	0.000
N of Valid Cases	539383		

4. Cluster analysis

In this study, we utilized departure time, departure date, and flight duration as features and employed the K-means algorithm to categorize U.S. airline companies. The selection of an appropriate K value, representing the number of clusters, was a crucial step. We carefully determined this value and proceeded with iterative executions of the K-means algorithm until convergence was achieved, ultimately yielding the final clustering results.

4.1. Cluster Analysis of Results

Based on the scale of this dataset, we chose the value of K to be 3. In Table 16, we use the k-means algorithm and observed distinct characteristic differences among the different clusters. The three clusters share the same departure dates and have similar flight durations. However, what sets them apart is that Cluster 2 has the latest departure times, Cluster 3 has the earliest departure times, and Cluster 1 falls in between the two.

Table 16. Final Cluster Centers

	Cluster		
	1	2	3
DayOfWeek	4	4	4
Time	801	1128	487
Length	130	131	136

4.2. Classification of U.S. Airlines Using Clustering Analysis Results

By conducting contingency table analysis shown in Table 17 and performing chi-squared test shown in Table 18, we found a significant two-tailed asymptotic significance level of 0.000, indicating a noteworthy relationship between airlines and the identified clusters. Consequently, we classified the airlines as follows: F9, FL, OO, US, YV, and B6 were categorized into Cluster 2, characterized by the latest departure times; AS, DL, UA, WN, and AA were placed into Cluster 3, characterized by the

DOI: 10.54254/2755-2721/55/20241522

earliest departure times; CO, EV, HA, MQ, OH, XE, and 9E were assigned to Cluster 1, with departure times falling between the two aforementioned clusters.

Table 17. Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1372.949[a]	34	0.000
Likelihood Ratio	1374.006	34	0.000
N of Valid Cases	539383		

Table 18. Airline * Cluster Number of Case Crosstabulation

		Cluster Number of Case			Total
		1	2	3	
Airline	9E	7431	6319	6936	20686
	AA	15607	14626	15423	45656
	AS	3387	3862	4222	11471
	B6	5769	6572	5771	18112
	CO	7370	6494	7254	21118
	DL	20607	19760	20573	60940
	EV	10449	8624	8910	27983
	F9	2053	2215	2188	6456
	FL	6875	7504	6448	20827
	HA	1975	1657	1946	5578
	MQ	13113	11179	12313	36605
	OH	4717	3776	4137	12630
	OO	17369	16712	16173	50254
	UA	8330	9176	10113	27619
	US	10991	11678	11831	34500
	WN	31917	30696	31484	94097
	XE	11557	9382	10187	31126
	YV	4731	5018	3976	13725
Total		184248	175250	179885	539383

4.3. Significance of Clustering

Combining the comparative analysis from the third section, we observed that the average departure time of flights not delayed was notably earlier than the delayed flights. Therefore, based on our classification of airlines, optimizing airlines within Cluster 2 holds substantial implications. These optimizations provide valuable guidance for developing more targeted aviation business strategies and streamlining operations.

DOI: 10.54254/2755-2721/55/20241522

5. Factor analysis

5.1. Adaptability Analysis

In Table 19, we conducted factor analysis on 'DayOfWeek', 'Time,' and 'Length.' The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was 0.499, indicating a moderate level of adequacy for the sample. Additionally, Bartlett's sphericity test yielded a significance value of 0.000, which is less than the significance level of 0.01, suggesting a significant relationship among the variables analyzed. This supports the suitability of performing factor analysis.

Table 19. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.499
Bartlett's Test of Sphericity	Approx. Chi-Square	327.320
	df	3
	Sig.	0.000

5.2. Common Factor Extraction

Table 20 shows that the initial eigenvalue of the first component is 1.024, greater than 1. The initial eigenvalue of the second component is 1.001, also greater than 1. The initial eigenvalues of the remaining components are less than 1. Therefore, selecting two common factors can achieve a cumulative contribution rate of 67.506%, indicating that these two common factors can explain approximately 67% of the total variance. This result is quite satisfactory.

Table 20. Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.024	34.134	34.134	1.024	34.134	34.134	1.020	34.004	34.004
2	1.001	33.372	67.506	1.001	33.372	67.506	1.005	33.502	67.506
3	0.975	32.494	100.000						

5.3. Factor Loadings

We applied the maximum variance method for factor rotation. In Table 21, we observed that the first common factor had substantial loadings on 'Length' and 'Time,' categorizing it as a spatiotemporal factor. The second common factor exhibited significant loadings on the 'DayOfWeek,' leading to its classification as the 'DayOfWeek' factor.

Table 21. Rotated Component Matrix

	Component	
	1	2
Time	0.772	
Length	-0.651	
DayOfWeek		0.918

5.4. Explained Variance by Common Factors

From the results shown in Table 22, it can be observed that the communalities for all three variables in the table exceed 0.5. This implies that more than 50% of the information from each original variable is

DOI: 10.54254/2755-2721/55/20241522

accounted for by the extracted common factors. Therefore, the extracted common factors effectively capture a significant portion of the information contained in the original variables.

Table 22. Communalities

	Initial	Extraction
DayOfWeek	1.000	0.843
Time	1.000	0.658
Length	1.000	0.524

6. Logistic regression

6.1. Logistic Regression Model Utility

The evaluation of logistic regression model aims to measure its accuracy, robustness, and reliability through appropriate evaluation metrics. In our study, Table 23 observed a prediction accuracy of 76.1% for the non-delayed flight group and 32.4% for the delayed flight group. Considering the entire sample, the overall prediction accuracy of the model was 56.6%. Further calculation yielded an F1 Score of approximately 0.398. Given the relatively weak correlation in the dataset, such results are deemed acceptable.

Table 23. Classification Table

Observed			Predicted		
			Delay		Percentage Correct
			0	1	
Step 1	Delay	0	227585	71534	76.1
		1	162318	77946	32.4
	Overall Percentage				56.6

6.2. Influence of Key Predictive Variables

In this section, we will delve into the utility of the logistic regression model, with a specific focus on exploring the correlation between independent variables and the occurrence of flight delays.

In Table 24, we have observed that later departure time ('Time') is associated with a higher likelihood of delays. Likewise, longer flight duration ('Length') tends to increase the probability of delays. Additionally, flights departing earlier in the week ('DayOfWeek') demonstrate a higher susceptibility to delays.

Table 24. Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	DayOfWeek	-0.029	0.001	400.229	1	0.000	0.971
	Time	0.001	0.000	12189.457	1	0.000	1.001
	Length	0.001	0.000	1060.569	1	0.000	1.001
	Constant	-1.175	0.012	10177.916	1	0.000	0.309

7. Conclusion

In this study, we applied data mining methods to investigate the factors influencing flight delays in the United States. Through comparative analysis, we conducted an in-depth exploration of the relationships between various factors and flight delays. Our findings revealed a significant association between

DOI: 10.54254/2755-2721/55/20241522

airlines and flight delays. Among delayed flights, Southwest Airlines (WN) had the highest proportion at 27.3%, while Delta Air Lines (DL) dominated among non-delayed flights, with a proportion of 11.2%. Additionally, a strong correlation was observed between the departure date and flight delays. Wednesdays saw the highest percentage of delayed flights at 17.6%, whereas Thursdays led among non-delayed flights at 16.8%. Moreover, flight delays were significantly related to flight duration and departure time. Delayed flights had an average flight duration of 129.66 minutes, slightly shorter than the average duration of 135.37 minutes for non-delayed flights. Similarly, the average departure time for delayed flights was 12:45, earlier than the average departure time of 14:09 for non-delayed flights. We also found that the origin and destination states of flights were significantly associated with flight delays. In delayed flights, the origin and destination states in California had the highest proportions, while Texas dominated for non-delayed flights. Utilizing cluster analysis, we categorized major U.S. airlines into three groups based on departure date, flight duration, and departure time differences. This classification provides valuable insights for optimizing the operational strategies of various airlines, particularly for the category of airlines departing late. Furthermore, factor analysis uncovered two critical factors, namely, a time-space factor and a departure date factor, which collectively explained 67.506% of the information contained in the data. Finally, logistic regression analysis revealed a positive correlation between departure time and flight delays. In other words, later departure times and longer flight durations increased the likelihood of flight delays. Additionally, an inverse relationship was found between departure dates and flight delays. Flights departing earlier in the week were more likely to be delayed.

This study's strengths lie in its comprehensive data analysis approach, providing a detailed exploration of factors contributing to flight delays. By considering multiple variables, we investigated airline categorization, highlighted the significance of various factors, and ensured the statistical significance and scientific rigor of our findings. These findings offer essential guidance for crafting precise aviation operational strategies and avoiding delays. By analyzing the departure states, destination states, departure time, departure date, flight duration, and airlines across different dimensions, we offer a holistic perspective on the multifaceted causes of flight delays. These highlights underscore the depth and breadth of this research and its potential impact on the aviation industry.

References

[1] P Belobaba, A Odoni and C. Barnhart, The global airline industry, 2019.
[2] https://data.worldbank.org/indicator/IS.AIR.PSGR?locations=US
[3] Song C, Guo J, Zhuang J. Analyzing passengers' emotions following flight delays-a 2011–2019 case study on SKYTRAX comments[J]. Journal of Air Transport Management, 2020, 89: 101903.
[4] Britto R, Dresner M, Voltes A. The impact of flight delays on passenger demand and consumer welfare[C]//Proceedings of the 12th World Conference on Transport Research, Lisbon. 2010.
[5] Victor C. The influence of flight delays on business travellers[D]. University of Pretoria, 2010.
[6] Ball M, Barnhart C, Dresner M, et al. Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the United States[J]. 2010.
[7] Anupkumar A. INVESTIGATING THE COSTS AND ECONOMIC IMPACT OF FLIGHT DELAYS IN THE AVIATION INDUSTRY AND THE POTENTIAL STRATEGIES FOR REDUCTION[J]. 2023.
[8] Kostiuk P F, Long D, Gaier E M. The economic impacts of air traffic congestion[J]. Air Traffic Control Quarterly, 1999, 7(2): 123-145.
[9] De Villemeur E, Ivaldi M, Quinet E, et al. The Social Cost of Air Traffic Delays[M]. Centre for Economic Policy Research, 2015.
[10] Kim, Young Jin, et al. "A deep learning approach to flight delay prediction." 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC). IEEE, 2016.
[11] Ding Y. Predicting flight delay based on multiple linear regression[C]//IOP conference series: Earth and environmental science. IOP Publishing, 2017, 81(1): 012198.

DOI: 10.54254/2755-2721/55/20241522

[12] Nigam R, Govinda K. Cloud based flight delay prediction using logistic regression[C]//2017 International Conference on Intelligent Sustainable Systems (ICISS). IEEE, 2017: 662-667.
[13] Manna S, Biswas S, Kundu R, et al. A statistical approach to predict flight delay using gradient boosted decision tree[C]//2017 International conference on computational intelligence in data science (ICCIDS). IEEE, 2017: 1-5.
[14] Chakrabarty, Navoneel, et al. "Flight arrival delay prediction using gradient boosting classifier." Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2. Springer Singapore, 2019.
[15] https://www.kaggle.com/datasets/jimschacko/airlines-dataset-to-predict-a-delay