REVA UNIVERSITY

Established as per the Section 2(f) of the UGC Act, 1956
Approved by AICTE, COA and BCI, New Delhi

3rd International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) – 2022

IEEE BANGALORE SECTION

IEEE REVA University Student Branch

Flight Delay Prediction for Indian Air Carriers with Explainable Artificial Intelligence

Jyoti Singh (Paper ID - 89)
REVA Academy for Corporate Excellence,
REVA University
jyoti.ba07@race.reva.edu.in
Rashmi Agarwal
Mithun Dolthody Jayaprakash

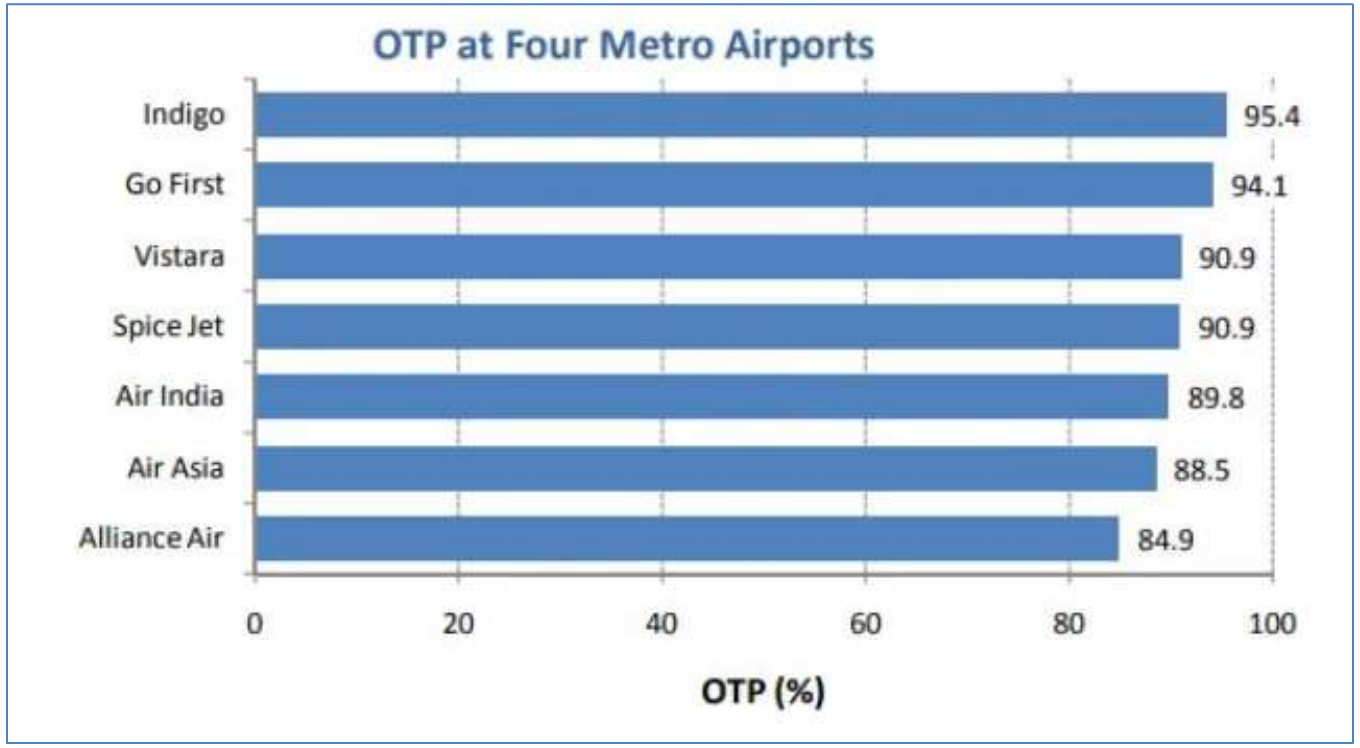
AGENDA

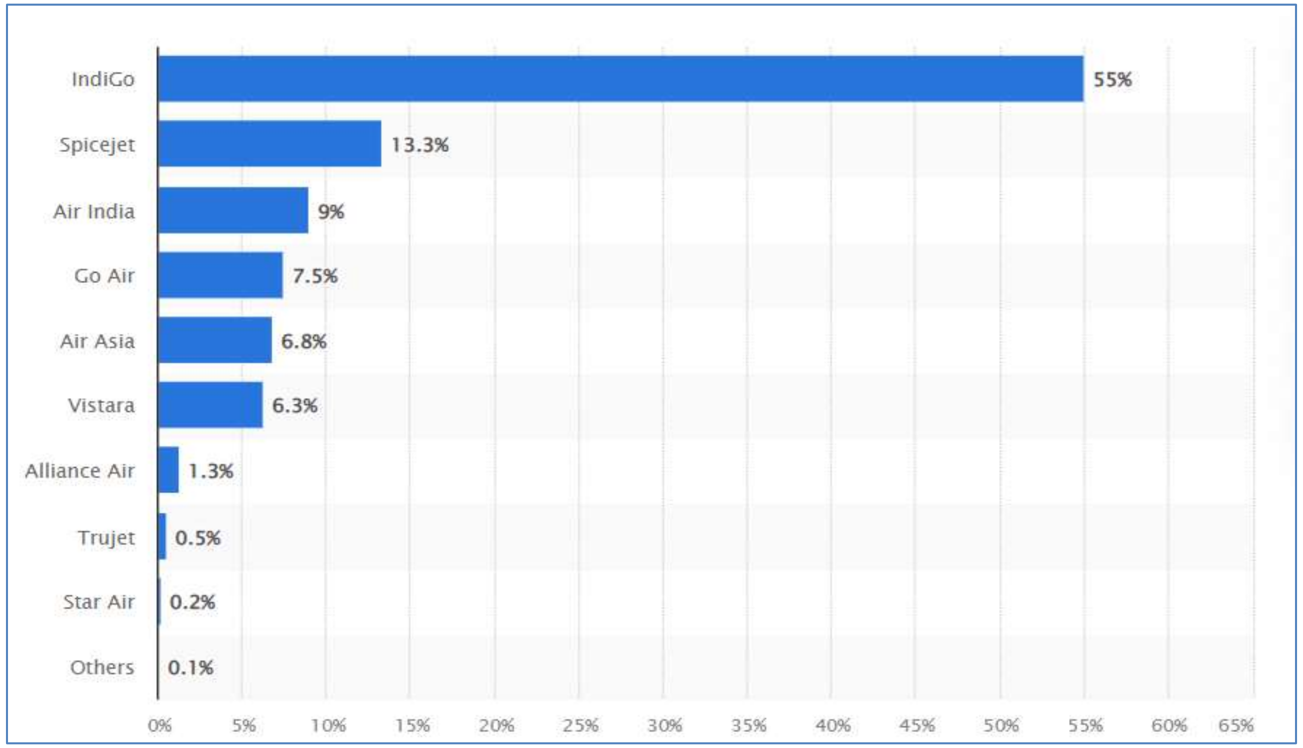
- Introduction
- Literature Review
- Problem Statement
- Objective of the research
- Methodology
- Business Understanding
- Data Understanding
- Data Preparation
- Modelling and Evaluation
- Conclusion
- Future Scope

INTRODUCTION

- A recent study reveals that India has become the third-largest civil aviation market in the world after US and China (according to International Air Transport Association - IATA).
- According to estimates, India will become one of the world's largest aviation markets by 2034 because of its visible growth trend.
- According to the Federal Aviation Administration (FAA), a flight is considered delayed if it is 15 minutes late than its scheduled arrival time.
- Indigo has achieved the highest on-time performance rating of 95.4% in the month of February 2022 by DGCA (Directorate General of Civil Aviation).
- This is the main reason IndiGo is the preferred low-cost airlines among flyers, resulting in the domestic aviation market share of approximately 55% in the financial year 2021-2022.

INTRODUCTION





* Image source: https://www.statista.com/statistics/575207/air-carrier-india-domestic-market-share/

INTRODUCTION

Impact of flight delays

- Passengers: Due to the lengthened travel durations and higher expenditures for lodging and meals, delays and cancellations cause travelers discomfort
- Airlines: Suffer penalties, fines, and additional operation costs, such as employing additional crews, accommodating interrupted passengers, flight re-positioning costs and aircraft retentions in airports. Delays have an impact on the airline's reputation as well
- Airports: The flight's delay has a negative influence on the airport's routine operations, and the subsequent flights are piled up as a result
- Environment: Due to increased fuel consumption and carbon emissions, aircraft delays have a significant negative impact on the environment

LITERATURE REVIEW

Paper Title	Journal	Objective
Research on Flight Delay Prediction Based on Random Forest [1]	IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), 2021	• The paper examines the distribution of delayed, on-time, and early arrivals flights by analyzing the departure flight data from Guangzhou Baiyun International Airport in June 2020 and selecting the data from ten landing airports. • It **analyses the significance of features such as the departure flight delay time, the scheduled flight time, the number of scheduled departure flights** on the day, the date, and the landing airport.
A Review on Flight Delay Prediction [2]	Transport Reviews , March 2017	• The paper presents a taxonomy and summarizes the **different approaches taken to predict flight delays**, with a particular focus on the use of machine learning • It also provides a history of important works that shows the connections between issues with flight delay prediction and the research trends to address them
Decreasing airline delay propagation by re-allocating scheduled slack [3]	IIE Transactions, 2010	• This study demonstrates how **delay propagation can be minimized by allocating planning slack**, and modifying the flight schedule just slightly while maintaining the original fleeting and crew scheduling choices.
Flight Delay Forecast due to Weather Using Data Mining [4]	Creative Commons AttributionNonCommercial-NoDerivatives 4.0 International License, 2015	• Authors **forecast airplane delays by studying the meteorological conditions** and using data mining techniques • WEKA and R were used to develop models for comparing the performance of different classifiers.

LITERATURE REVIEW

Paper Title	Journal	Objective
Prediction of weather-induced airline delays based on machine learning algorithms [5]	IEEE, 2016	• Research was to **mitigate the consequences of data imbalance** brought on by training data with weather attributes • For anticipating specific flight delays, they have employed methods including Decision Trees, AdaBoost, and K-Nearest Neighbors
Airline Delay Predictions using Supervised Machine Learning [6]	International Journal of Pure and Applied Mathematics, 2018	• The primary goal of this project is to **predict airline delays caused by various factors** • Performs predictive analysis, which includes a variety of statistical methods from **supervised machine learning and data mining**, which examines recent and previous data to create predictions assess about upcoming delays, utilizing Python's Regression Analysis with regularization technique

GAP

❑ Very minimal research papers focused on flight delays of Indian carriers

❑ This research primarily focuses on **forecasting flight delays and arrival time delays for Indian air carriers** in phases on the designated routes

❑ Additionally, the primary goal is to obtain practical insights on the **major factors causing flight delays for the most troublesome or delayed flights** using an explainable AI technique

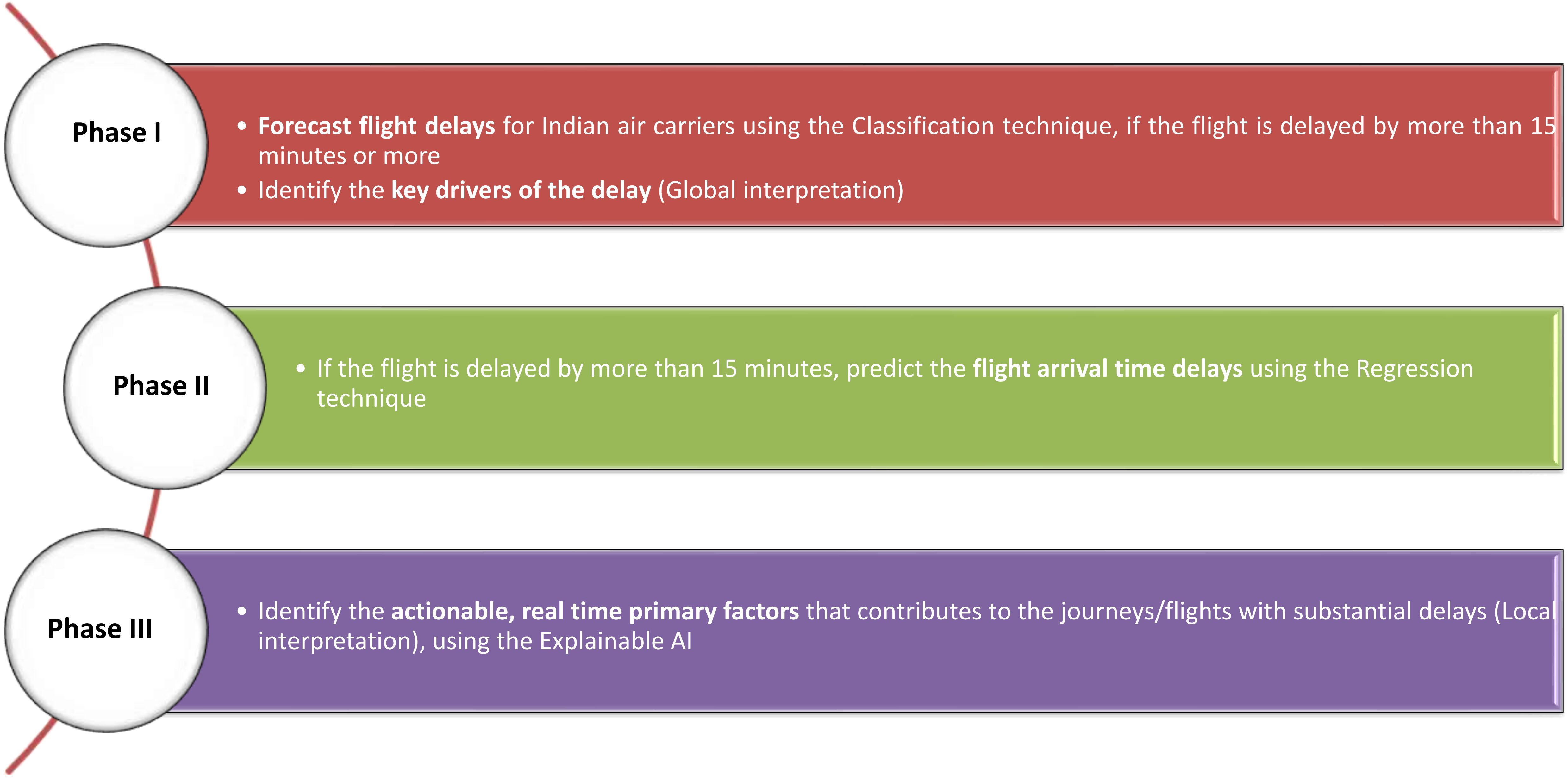
PROBLEM STATEMENT

- To predict the flight delays for Indian air carriers based on various controllable and uncontrollable factors

- To identify the primary causes of aircraft delays

- To identify the key drivers of delay for the flights with substantial delays using Explainable AI

OBJECTIVES OF THE STUDY



Phase I

- **Forecast flight delays** for Indian air carriers using the Classification technique, if the flight is delayed by more than 15 minutes or more
- Identify the **key drivers of the delay** (Global interpretation)

Phase II

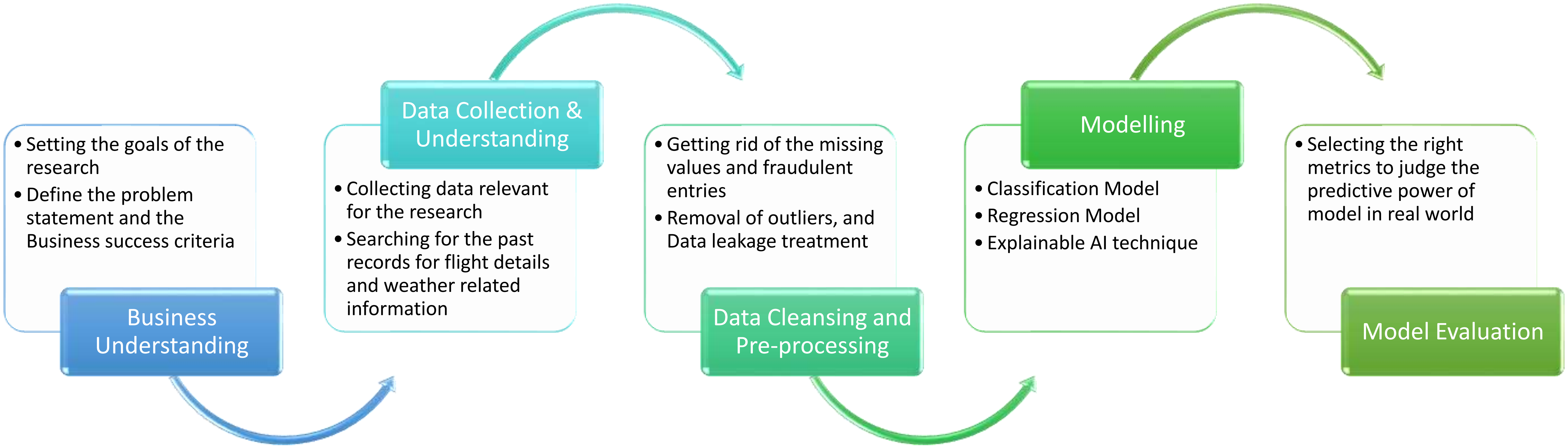
- If the flight is delayed by more than 15 minutes, predict the **flight arrival time delays** using the Regression technique

Phase III

- Identify the **actionable, real time primary factors** that contributes to the journeys/flights with substantial delays (Local interpretation), using the Explainable AI

METHODOLOGY

CRISP-DM



Business Understanding

- Setting the goals of the research
- Define the problem statement and the Business success criteria

Data Collection & Understanding

- Collecting data relevant for the research
- Searching for the past records for flight details and weather related information

Data Cleansing and Pre-processing

- Getting rid of the missing values and fraudulent entries
- Removal of outliers, and Data leakage treatment

Modelling

- Classification Model
- Regression Model
- Explainable AI technique

Model Evaluation

- Selecting the right metrics to judge the predictive power of model in real world

BUSINESS UNDERSTANDING

- The primary goal of the research is to forecast flight delays and key drivers affecting the delays for the selected airports and airlines
- The important element of the business objective is also to identify the primary causes of delay for the specific trips with substantial delay
- 5 airports chosen for the study are Delhi, Mumbai, Bangalore, Hyderabad, and Kolkata as they contribute more than half of all domestic passenger flow and it is also represents the geographical diversity of the country.
- Indigo, Air India, SpiceJet, GoAir, Air Asia, and Vistara; the six airlines under consideration, collectively control 99.4% of the domestic market. For this Airline trends, ratings and Market share are considered for the research
- The Federal Aviation Administration (FAA) estimates that weather delays account for about 70% of aircraft delays. Consequently, the research considers weather considerations as well to anticipate delays

DATA UNDERSTANDING

- The flight information dataset consists of 14951 observations and 22 features during a period of two years, from January 2018 to January 2020
- The weather dataset includes 14 meteorological features and 90600 observations spread throughout two years, from January 2018 to January 2020
- Airport scores (Departure and Arrival)
  - Airport Rating – Overall rating of the airport based on the operations, management, and development of the airport
  - Airport On-Time Performance Rating – Airport performance based on the aircraft arrival at the gate under 15 minutes of the scheduled arrival time
  - Airport Service Rating - Ratings from the customer reviews, which take into account the cleanliness of the airport, wait times, airport shopping, cafes and restaurants, wi-fi availability, and staff services
- Airlines Ratings
  - Carrier Rating – Based on the quality of the airline's overall services
  - Carrier Market Share – Airline market share based on their market position, Revenue generation, and the Operating Cost structure
  - Carrier Load Factor – Measures the proportion of seats that have been occupied by passengers
  - Carrier On-Time Performance rating - Performance of the carrier as measured by the aircraft arriving at the gate within 15 minutes of the planned arrival time

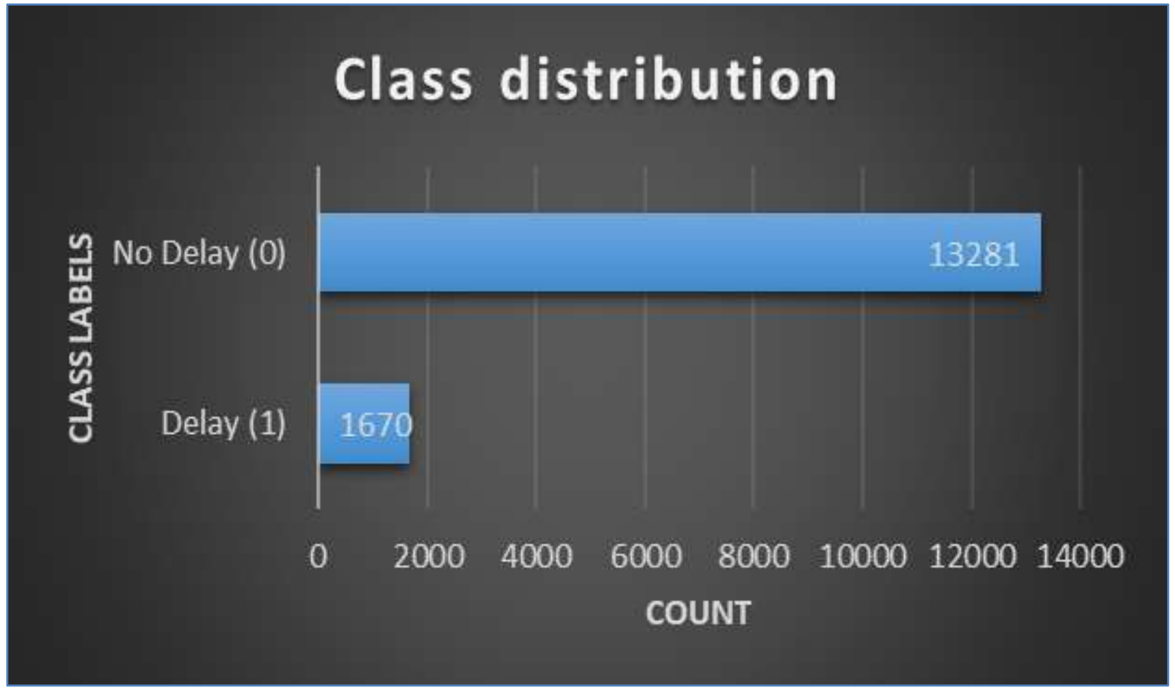
DATA UNDERSTANDING

- Flight Departure and Arrival details
  - Departure and Arrival Airports
  - Scheduled Departure and Arrival time
  - Duration of the flight
  - Actual Departure and Arrival time
- Weather details (Departure and Arrival Airport)
  - Dew Point
  - Wind Gust, Wind Speed, and Wind direction
  - Cloud Cover
  - Humidity and Precipitation
  - Pressure and Temperature
  - Visibility

DATA UNDERSTANDING

Exploratory Data Analysis

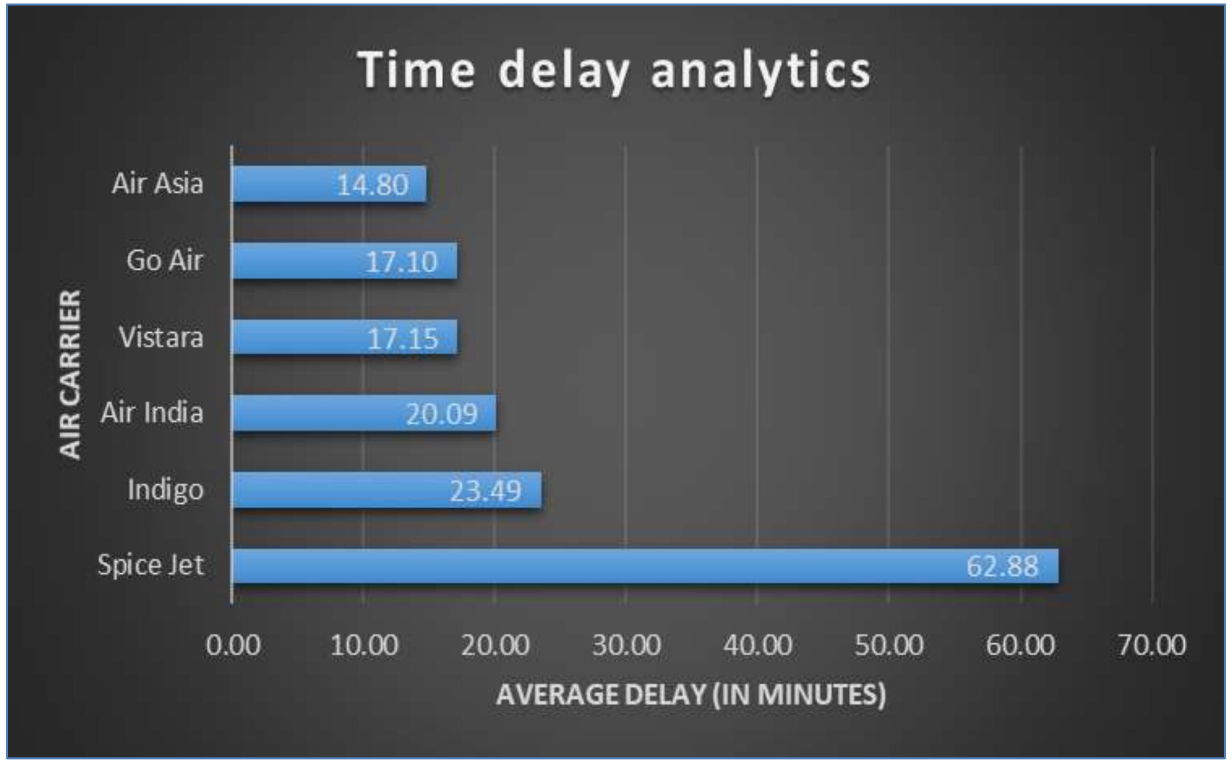
- The chart represents the class distribution of Delayed v/s Non-delayed flights. We can infer from the chart that there is a severe imbalance in the classes, with the No-delay class accounting for 88% of the population data



DATA UNDERSTANDING

Exploratory Data Analysis

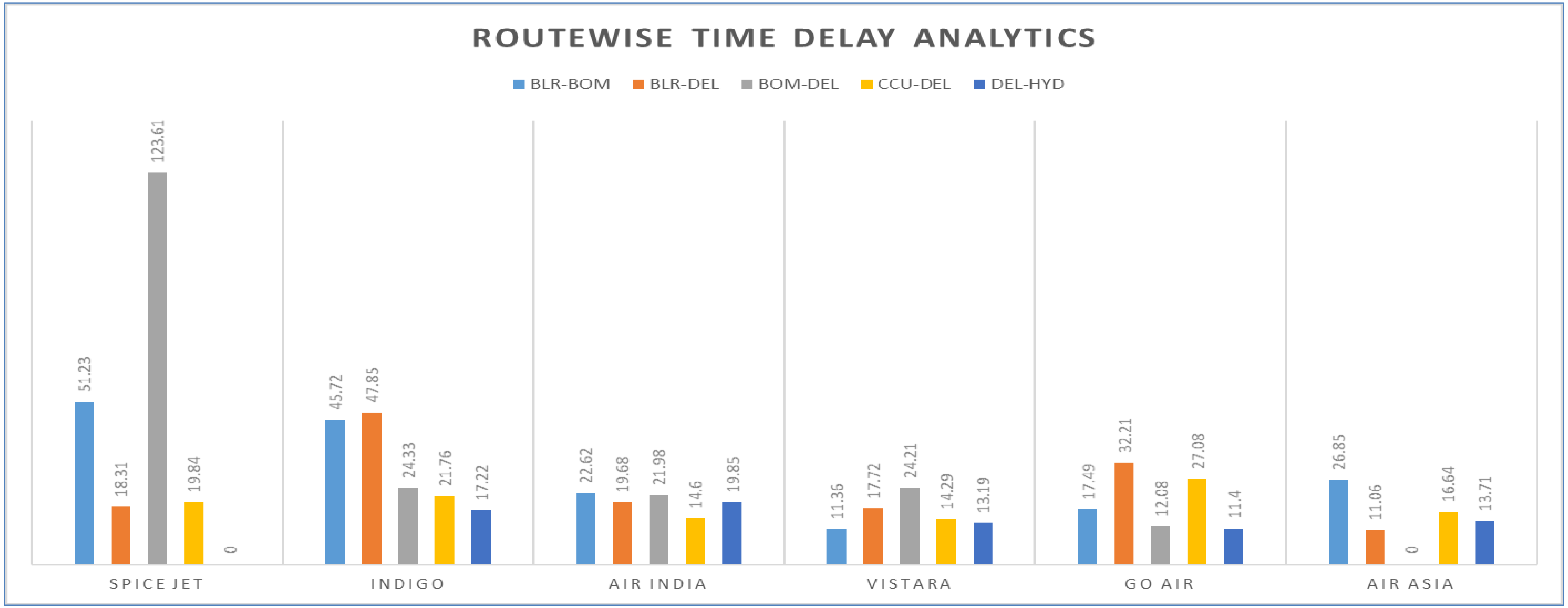
- The chart represents the Average Delay time (in minutes) for various carriers. Having an average delay of 62 minutes, Spice Jet is identified as the airline with the greatest frequency of delays. The least delayed airline is identified to be Air Asia, with an average delay of just 14.80 minutes



DATA UNDERSTANDING

Exploratory Data Analysis

- The chart represents the Route-wise Time delay analytics. It can be inferred that SpiceJet tops the list of airlines with the most delays on the BLR-BOM and BOM-DEL routes



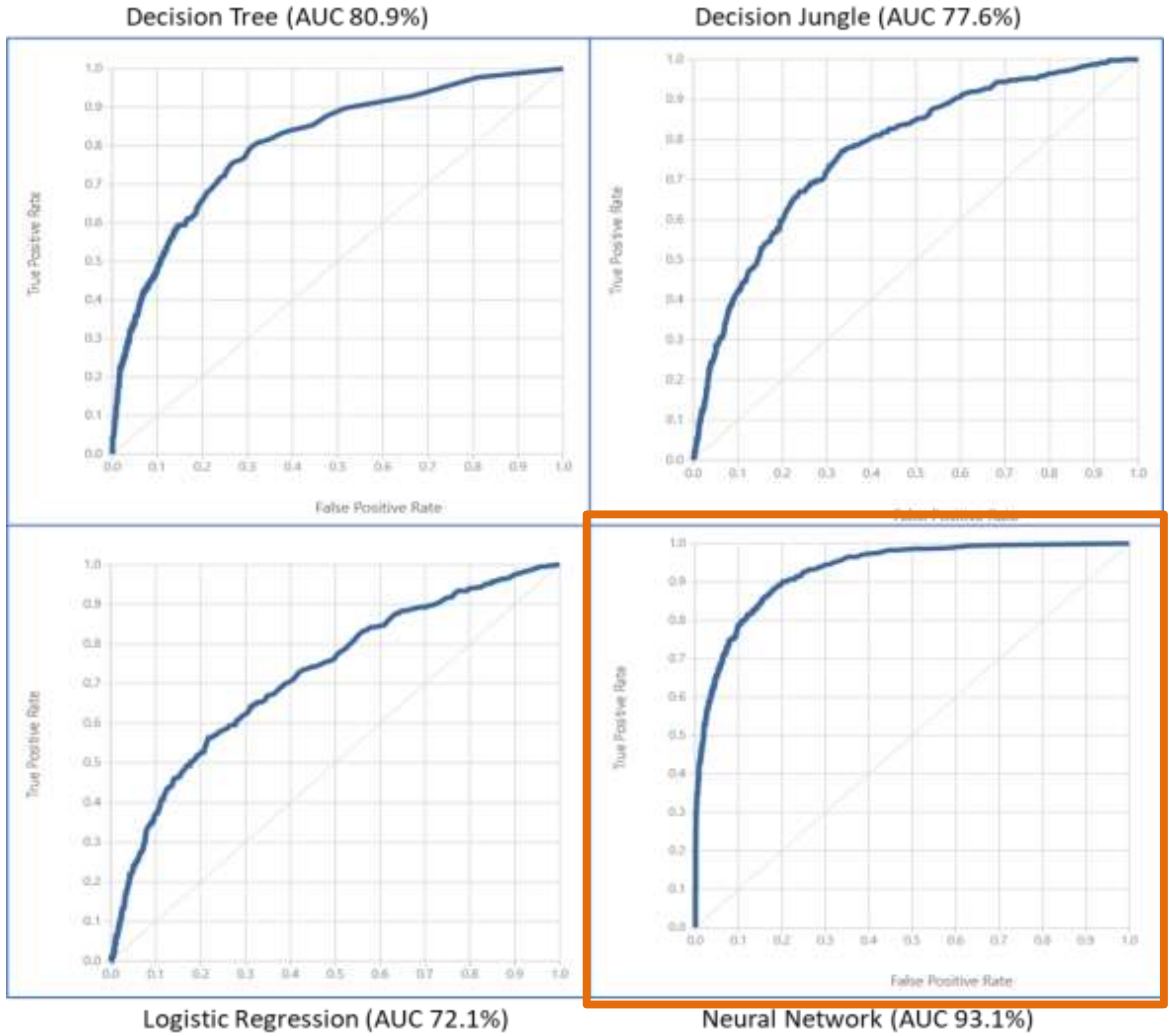
DATA PREPARATION

- Data Leakage Treatment
- Conversion of Categorical Column to Numerical column
- Outlier Removal: Outliers in the arrival delay are removed using interquartile range approach. In the current dataset, the upper whisker is 34 minutes and the lower whisker is -46 minutes
- Data Normalization: The research uses MinMax scaling technique to normalize the dataset between range 0 and 1
- Class Imbalance handling: Undersampling of the majority class i.e. class 0, is the resampling approach employed in the research to address the class imbalance. Post undersampling the class sizes were divided in 40: 60 ratio

MODELLING AND EVALUATION

Phase 1 – Classification Techniques

- Figure depicts the ROC (Receiver Operating Characteristic) curve of all the Classification models with Area under the ROC curve (AUC). Neural Network Classifier tops the chart with an AUC of 93.1%.



MODELLING AND EVALUATION

Phase 1 – Classification Techniques

- The table indicates that the Neural Network classifier is the best non-linear model for the dataset under consideration, with an accuracy of 92.5% and a precision of 73.6%

Model	Key Metrics			
	Accuracy	Precision	Recall	F1 Score
Decision Tree	68.50%	26.30%	43.40%	32.75%
Two-class Decision Jungle	88.9%	40.7%	6.8%	11.7%
Logistic Regression	88%	0	0	0
Neural Network classifier	92.5%	73.6%	54.6%	62.7%

MODELLING AND EVALUATION

Phase 1 - Top 5 important variables

- The top 5 independent variables influencing flight delays are listed in Table II, with flight Duration topping the list with 13.2% followed by Arrival Airport On-Time Rating with 5.9%.

Sr. No.	Independent variable	Importance
1	Duration	13.2%
2	Carrier_OnTimeRating	5.9%
3	Deprture_AirportOnTimeRating	5.6%
4	Arrival_AirportServiceRating	5.4%
5	Departure_windspeedKmph	4.9%

MODELLING AND EVALUATION

Phase 2 – Regression Techniques

- With the lowest RMSE and MAE values and maximum co-efficient of determination, we can conclude that the Linear Regression model with principal components is the best model for predicting delay in flight arrival time

Model	**Key Metrics**		
	Root Mean Squared Error (RMSE)	***Mean Absolute Error (MAE)***	***Co-efficient of Determination (R^2)***
Linear Regression	121.03	40.64	3.43
Regression Neural Network (MLP Regressor)	97.81	7.51	48.91
Linear Regression with PCA	3.52	2.20	93.44

MODELLING AND EVALUATION

Phase 2 – Principal Component Analysis

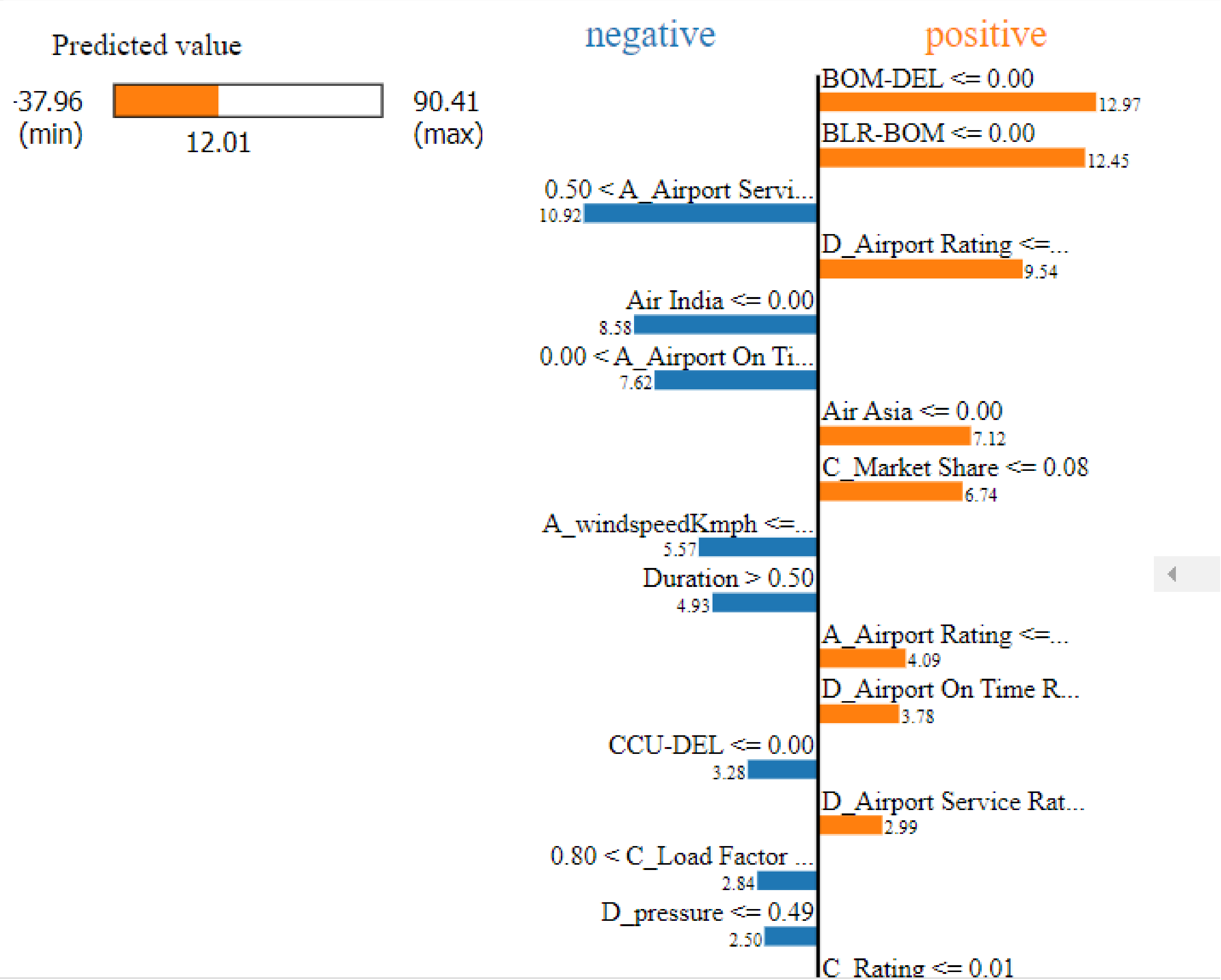
- In the current research, 14 principal components are created using the PCA technique
- These combinations are made in a way that most of the information included in the original variables is condensed into the first components, which are the new variables (i.e., principal components), and these are uncorrelated

Features	**Factors Significance**		
	Factor 8 (Carrier) 0.663	**Factor 10 (Wind) 0.643**	**Factor 6 (Mistiness) 0.543**
Carrier Rating	0.975		
Vistara	0.954		
Departure Wind Gust		0.88	
Departure Wind speed		0.866	
Departure Cloud cover			0.701
Arrival Dew point			0.631
Departure Humidity			0.614

MODELLING AND EVALUATION

Phase 3 – Explainable AI (XAI)

- Local Interpretable Model-agnostic Explanations (LIME) package used to provide local explanations for SpiceJet flight from BLR-DEL



MODELLING AND EVALUATION

Phase 3 – Explainable AI (XAI)

- The features that can be controlled and tackled easily are the Airport Rating, On-Time Performance rating of the Airport, and Airline
- Although weather parameters cannot be controlled, Airline can plan to schedule the flight time based on the weather conditions

Sr. No.	Controllable Features	Uncontrollable features
1	Departure Airport Rating	Route
2	Carrier Market share	Departure precipitation
3	Arrival Airport Rating	Arrival visibility
4	Departure Airport On-Time Performance rating	Departure Cloud cover
5	Carrier On-Time Performance Rating	Arrival precipitation

CONCLUSION

- The Neural Network Classifier model emerged as the best model with an accuracy of 92.5% and a precision of 73.6% for predicting the delayed flight.
- Linear Regression Model with Principal Components is identified as the best Regression model with the lowest RMSE of 3.52, MAE of 2.20, and maximum coefficient of determination(R square) of 93.44 for predicting delay in flight arrival time.
- The XAI technique can be used by the airport and airline authorities to locally interpret the main causes of the arrival delays for the trip with substantial delays.
- Market share, carrier OTP rating, load factor of the airline, OTP ratings of the departure and arrival airports are identified as the most significant controllable features affecting flight delays.
- The destination Airport’s wind speed, precipitation, humidity, and visibility are identified as the most significant uncontrollable features affecting flight delays.

FUTURE SCOPE

- In-depth analysis of the factors influencing airports' and airlines' OTP ratings.
- The research's forecasts and key driver analyses can be used to develop more real-time applications for passengers and authorities.
- A broader dataset can be taken into consideration for accurately capturing the impact of controllable and uncontrollable factors
- The research scope can be extended to other countries. One can broaden the research's reach by including flight information from international flights rather than simply domestic ones
- The research can be extended to create a prediction model for the departure time delay, as passengers tend to complain more about uncertainty and departure delays

Thank you

APPENDIX

REFERENCES

[1] P. Hu, J. Zhang and N. Li, 2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT)

[2] Alice Sternberg, Jorge Soares, Diego Carvalho, Eduardo Ogasawara, Transport Reviews , March 2017

[3] Shervin AhmadBeygi, Amy Cohn and Marcial Lapp, IIE Transactions, 2010

[4] Adrian Alexander, Arteche Simmons, Creative Commons AttributionNonCommercial-NoDerivatives 4.0 International License, 2015

[5] S. Choi, Y. J. Kim, S. Briceno and D. Mavris, IEEE, 2016

[6] PranalliChandraa, Prabakaran.N and Kannadasan.R, International Journal of Pure and Applied Mathematics, 2018