A PROJECT REPORT

on

“FLIGHT FARE PREDICTION ”

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR’S DEGREE IN
INFORMATION TECHNOLOGY

BY

SARTHAK AGARWAL

UNDER THE GUIDANCE OF
MAYANK GUPTA SIR
Senior Software Engineer

School of Computer Engineering

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024
November 2024

A PROJECT REPORT
on

"FLIGHT FARE PREDICTION"

Submitted to
KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR'S DEGREE IN
INFORMATION TECHNOLOGY

BY

SARTHAK AGARWAL 2106319

UNDER THE GUIDANCE OF
MR. MAYANK GUPTA

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA -751024
November 2024

KIIT Deemed to be University
School of Computer Engineering
Bhubaneswar, ODISHA-751024

CERTIFICATE

This is certify that the project entitled

“FLIGHT FARE PREDICTION”

Submitted by

SARTHAK AGARWAL 2106319

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering in Information Technology at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2024-2025, under our guidance.

Date: 12/11/2024

Mr. Mayank Gupta
Project Guide

Acknowledgements

I would like to express my sincere gratitude to all those who contributed to the successful completion of this project. This achievement would not have been possible without the support and guidance of many individuals and teams.

Firstly, I would like to thank InterGlobe Technology Quotient for providing me with the opportunity to work on this project as an intern. The resources and environment offered by the company were instrumental in achieving the project goals within the time frame.

I am deeply grateful to my general manager, Rajeev Kumar Yadav Sir, for their continuous support, valuable feedback, and insightful guidance throughout the project. Their leadership and encouragement were essential in steering the project to its successful completion.

I would also like to extend my thanks to the entire Software Developer team. Your dedication, hard work, and collaborative spirit made this project a rewarding experience. Special thanks to Mayank Gupta Sir for going above and beyond in your efforts, ensuring that all aspects of the project were handled with the utmost professionalism and care.

I would also like to thank Dishank Sir and Ranjeet Sir for their technical support and for providing the necessary tools and infrastructure to bring this project to life.

Thank you all for your unwavering support and dedication.

Sarthak

Sarthak Agarwal

ABSTRACT

Air travelers (the buyers) usually look for the best time of year to purchase flights in order to save as much money as possible, while airlines (the sellers) always try to improve their revenue by altering the pricing for the same service. Based on all the pertinent information, including past sales, market demand, customer profile, and behavior, the sellers can decide whether to raise or lower tickets at different times before departure dates. On the other hand, customers have little access to information that would assist them determine whether to purchase a flight right now or postpone it. In this work, we propose a novel approach that could help the buyer predict price changes even when there are no official airlines.

This project aims to predict flight fares using machine learning techniques. The dataset, sourced from Kaggle, was preprocessed to handle missing values and extract relevant features. Exploratory Data Analysis (EDA) was conducted to gain insights into the data distribution and relationships between variables. Feature engineering techniques were employed to create new informative features. Machine learning models, including Linear Regression, Random Forest Regressor, and Extra Trees Regressor, were trained and evaluated. The Extra Trees Regressor emerged as the top-performing model, achieving a significant R-squared score on the test set. Future work could involve exploring advanced techniques like deep learning and incorporating real-time data for more accurate predictions.

Keywords: Machine Learning, Feature Engineering, Linear Regression, Random Forest, Extra Trees Regressor, Data Mining, Exploratory Data Analysis.

Contents

Chapter 1

Introduction

The airline industry is a highly competitive market where prices can fluctuate rapidly due to various factors such as demand, supply, and external events. As a result, predicting flight fares has become a crucial task for airlines, travel agencies, and passengers alike. Accurate fare predictions can help airlines optimize their pricing strategies, while passengers can benefit from making informed decisions about their travel plans. This project aims to develop a flight fare prediction model using machine learning algorithms to provide accurate and reliable fare predictions.

The importance of this project lies in its potential to provide valuable insights into the airline industry's pricing dynamics. By analyzing historical data and identifying key factors that influence fare prices, this project can help stakeholders make data-driven decisions. Moreover, the development of a robust fare prediction model can also contribute to the advancement of the field of machine learning and its applications in real-world problems.

The report includes visualizations and tables to illustrate the results and facilitate understanding. Overall, this project aims to contribute to the development of a reliable and accurate flight fare prediction model that can benefit the airline industry and its stakeholders.

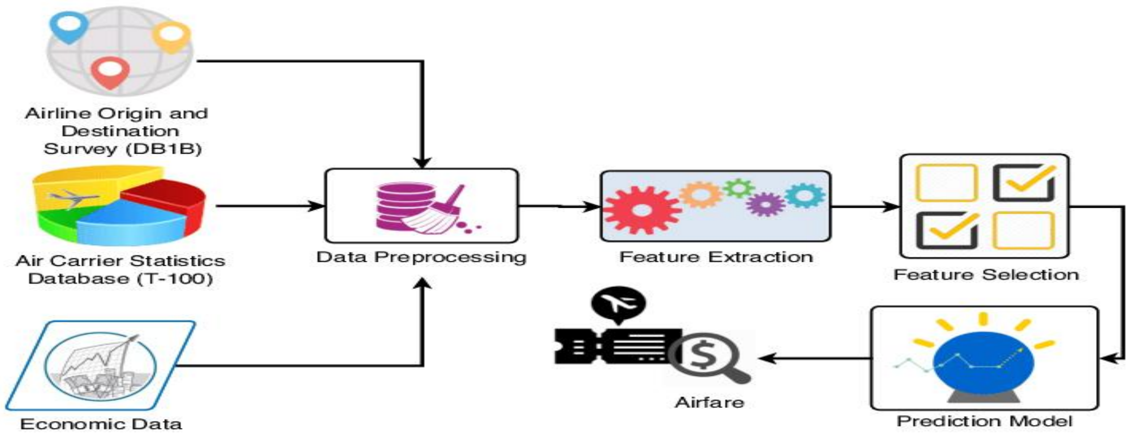


FIG 1: A brief overview of what the model works like.

Chapter 2

Basic Concepts

This section provides an overview of the basic concepts related to the tools and techniques used in this project.

2.1 Machine Learning:

Machine learning is a subset of artificial intelligence that involves training algorithms to learn from data and make predictions or decisions without being explicitly programmed. In this project, machine learning algorithms are used to predict flight fares based on historical data. The goal of machine learning is to develop a model that can generalize well to new, unseen data and make accurate predictions.

2.2 Supervised Learning:

Supervised learning is a type of machine learning where the algorithm is trained on labeled data, meaning the data is already tagged with the correct output. In this project, supervised learning is used to train the model to predict flight fares based on input features such as departure city, arrival city, airline, and travel dates. The model learns to map the input features to the corresponding output (fare) based on the labeled data.

2.3 Regression Analysis:

Regression analysis is a statistical technique used to establish a relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). In this project, regression analysis is used to predict continuous values (flight fares) based on multiple input features. The goal of regression analysis is to develop a model that can accurately predict the target variable based on the input features.

2.4 Linear Regression:

Linear regression is a type of regression analysis where the relationship between the dependent variable and independent variables is modeled using a linear equation. In this project, linear regression is used as a baseline model to predict flight fares. Linear regression assumes a linear relationship between the input features and the target variable, which may not always be the case in real-world scenarios.

2.5 Random Forest:

Random forest is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy and robustness of predictions. In this project, random forest is used to predict flight fares based on input features. Random forest works by training multiple decision trees on different subsets of the data and then combining their predictions to produce a final output.

2.6 Feature Engineering:

Feature engineering is the process of selecting and transforming raw data into features that are more suitable for modeling. In this project, feature engineering is used to extract relevant features from the raw data, such as departure city, arrival city, airline, and travel dates. Feature engineering is a critical step in machine learning, as the quality of the features can significantly impact the performance of the model.

2.7 Data Pre-processing:

Data pre-processing is the process of cleaning, transforming, and preparing raw data for modeling. In this project, data pre-processing is used to handle missing values, outliers, and data normalization. Data pre-processing is an essential step in machine learning, as it can significantly impact the performance and accuracy of the model.

2.8 Evaluation Metrics:

Evaluation metrics are used to measure the performance of machine learning models. In this project, evaluation metrics such as R-squared score are used to evaluate the performance of the models. These metrics provide insights into the accuracy and robustness of the models, and help to identify areas for improvement.

2.9 Extra Trees Regressor:

An Extra Trees Regressor is an ensemble learning algorithm that combines multiple decision trees to improve the accuracy and robustness of predictions. It is similar to a Random Forest Regressor, but with a few key differences.

In an Extra Trees Regressor, each decision tree is trained on the entire dataset, rather than a random subset of features. This can lead to improved performance, especially when dealing with datasets that have a large number of features.

Additionally, Extra Trees Regressor uses a technique called “extremely randomized trees”, which involves randomly selecting a feature and a split point for each node in the tree. This can help to reduce over-fitting and improve the robustness of the model.

Chapter 3

Problem Statement & Requirement Specifications

How can we develop a predictive model that accurately forecasts flight fares, enabling customers to make informed decisions and find the best deals, while also helping airlines to optimize their pricing strategies and improve revenue management?

3.1 Project Planning:

The following steps can be followed while planning to execute the project development:

Step 1: Define Project Scope

- Predict flight fares using historical data.
- Develop a machine learning model to forecast fares.
- Evaluate the performance of the model using metrics such as R-squared score.

Step 2: Gather Requirements

- Collect historical flight fare data.
- Identify relevant features to be used in the model, such as:
  - Departure and arrival cities
  - Departure and arrival dates
  - Airlines and flight numbers
- Determine the type of machine learning algorithm to be used (e.g. linear regression and random forests)

Step 3: Identify Risks and Assumptions

- Risk: Insufficient data or poor data quality.
- Assumption: Historical data is representative of future trends.
- Mitigation strategy: Collect more data or use data augmentation techniques.

Step 4: Establish Communication Plan

- Regular meetings with mentors to discuss progress and challenges.
- Update the mentors on project milestones.

3.2 Project Analysis:

After the requirements are collected or the problem statements is conceptualized, the following steps were taken to analyze the project:

Step 1: Review and Refine Requirements

- Review the requirements document to ensure it is complete and accurate
- Refine the requirements to remove any ambiguity or inconsistencies
- Ensure that the requirements are measurable, achievable, relevant, and time-bound (SMART)

Step 2: Identify Ambiguities and Gaps

- Identify any ambiguities or gaps in the requirements, such as:
- How to handle missing or erroneous data
- How to account for seasonal and holiday fluctuations
- How to evaluate the performance of the model
- Clarify any unclear or incomplete requirements

Step 3: Analyze Technical Feasibility

- Evaluate the technical feasibility of the project, including:
- Availability of historical data
- Computational resources required for model training and deployment
- Integration with existing systems or infrastructure

Step 4: Identify Data Requirements

- Identify the data requirements for the project, including:
- Historical flight fare data
- Relevant features and variables
- Data formats and storage requirements

Step 5: Develop a High-Level Design

- Develop a high-level design for the project, including:
- Data pre-processing and feature engineering pipeline
- Machine learning model architecture
- Deployment strategy and infrastructure requirements

3.3 System Design:

3.3.1 Design Constraints:

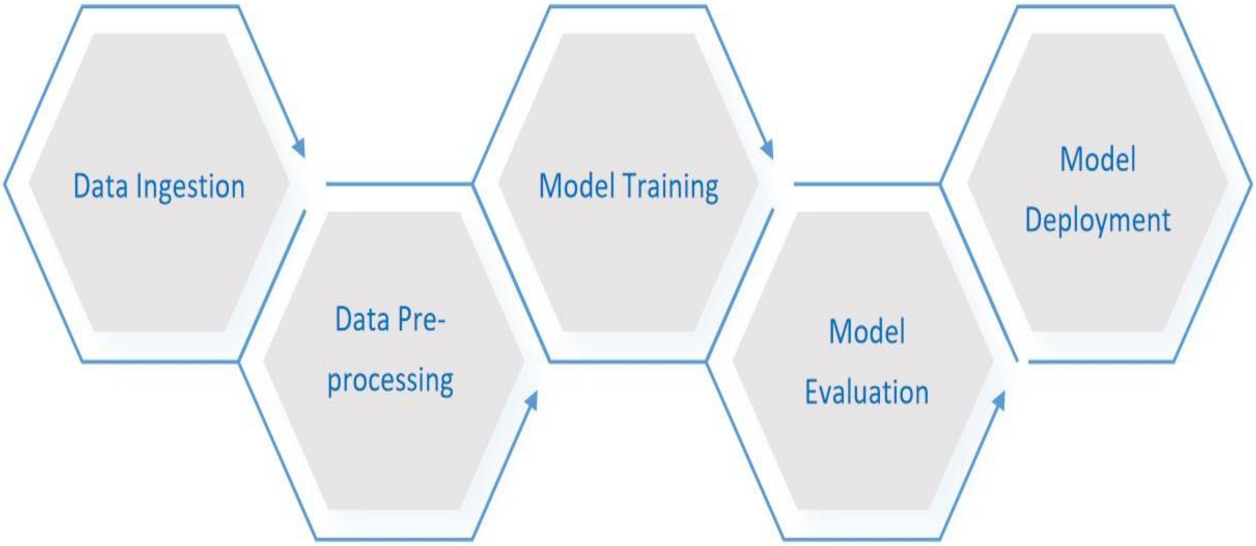
- The system is designed to run on any Integrated Development Environment, with Python 3 as the programming language.
- The system uses various libraries and frameworks, including Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and XGBoost.

3.3.2 System Architecture & Block Diagram:

The system architecture consists of the following components:

- Data Ingestion: Historical flight fare data is collected and ingested into the system.
- Data Pre-processing: The ingested data is cleaned, transformed, and feature-engineered to prepare it for model training.
- Model Training: The pre-processed data is used to train a machine learning model, such as a random forest.
- Model Evaluation: The trained model is evaluated using metrics such as R-squared score.
- Model Deployment: The trained model is deployed in a user-friendly interface, allowing users to input parameters and receive predicted flight fares.

The block diagram for the system architecture is as follows:



Chapter 4

Implementation

In this section, we present the implementation done during project development.

4.1 Methodology:

In this project, we aimed to develop a predictive model that can accurately forecast flight fares based on various factors such as airline, class, stops, departure time, arrival time, source city, destination city, and days left for departure. To achieve this, we employed a range of machine learning algorithms and techniques.

The methodology adopted for this project can be summarized as follows:

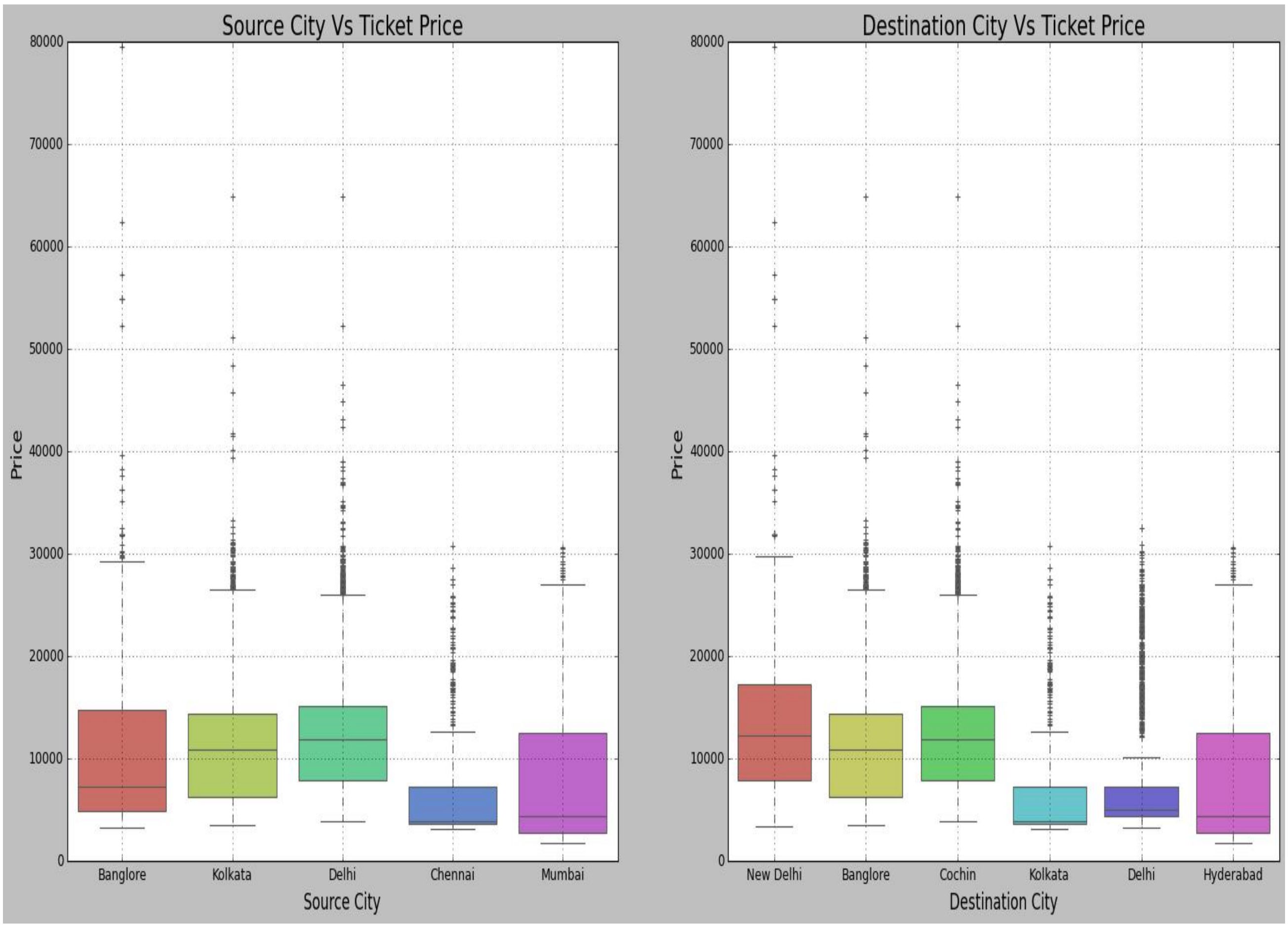
1. Data Pre-processing: We collected a dataset of flight fares and preprocessed it by handling missing values, encoding categorical variables, and scaling numerical variables.
2. Exploratory Data Analysis: We performed EDA to understand the distribution of variables, identify correlations, and visualize relationships between variables.
3. Feature Engineering: We extracted relevant features from the dataset, including flight duration, days left for departure, and class.
4. Model Selection: We selected a range of machine learning algorithms, including Linear Regression, Random Forest Regressor, Extra Trees Regressor, Lasso Regression.
5. Model Evaluation: We evaluated the performance of each model using metrics such as R2 Score.
6. Hyper-parameter Tuning: We tuned the hyper-parameters of the best-performing models to optimize their performance.

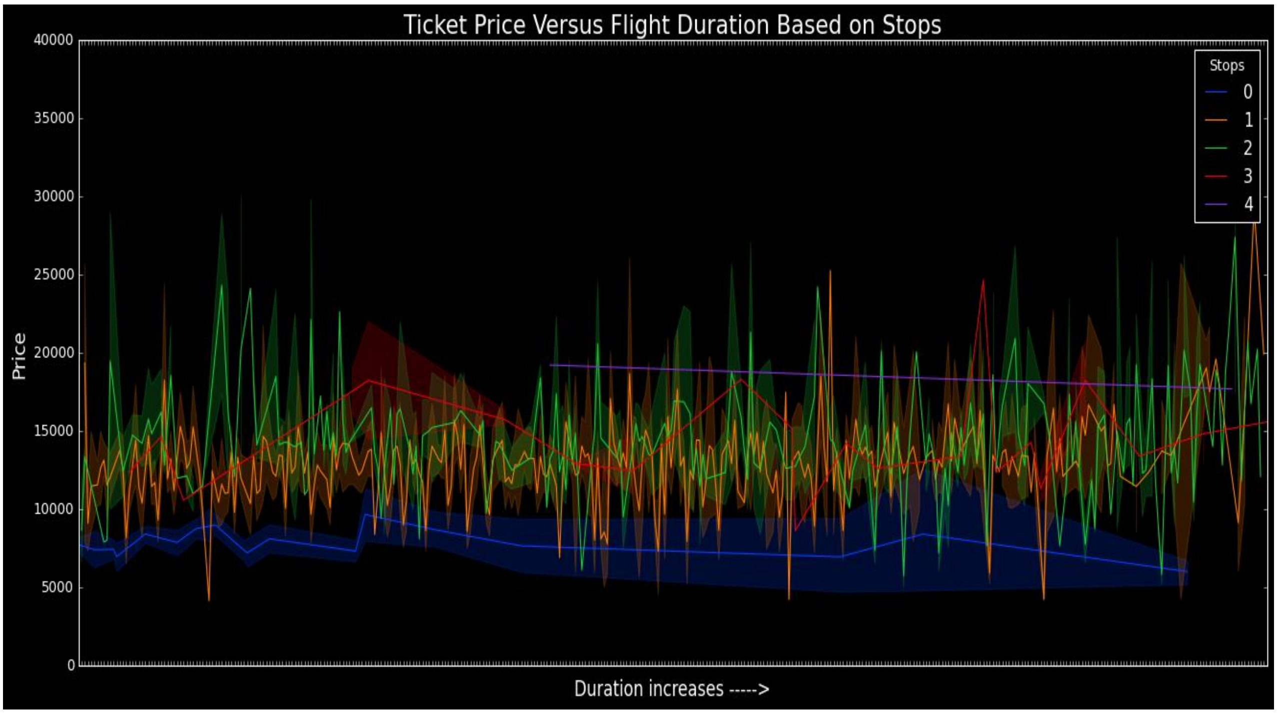
4.2 Testing Plan:

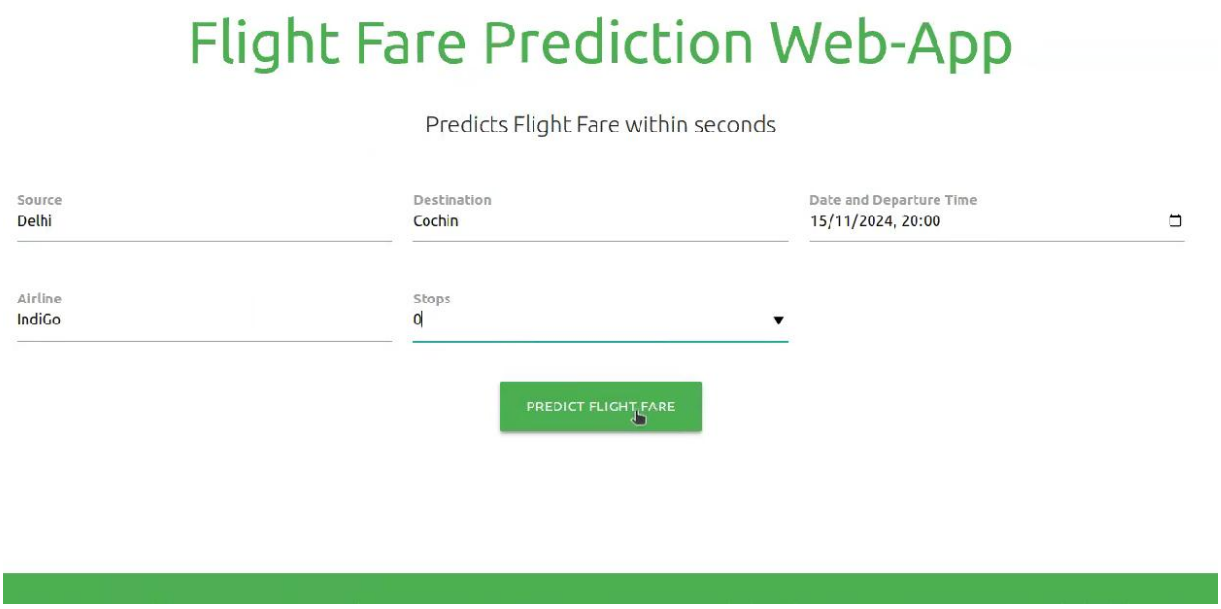
To verify the performance of our models, we created a testing plan with the following tests:

ID	Test Name	Test Condition	System Behavior	Expected Result
T01	Same city	Entered same city in Source and Destination.	Source and Destination can't be same. Please try again.	Result was as expected.
T02	Date and Departure Time	Date and Departure Time are not entered.	Please fill in this field. (Date and Departure Time in this case)	Result was as expected.
T03	Airline Test	Airline was not entered.	Please fill in this field. (Airline in this case)	Result was as expected.

4.3 Results Analysis:






Flight Fare Prediction Web-App

Predicts Flight Fare within seconds

Source
Delhi

Destination
Cochin

Date and Departure Time
15/11/2024, 20:00

Airline
IndiGo

Stops
0

Chapter 5

Standards Adopted

5.1 Design Standards:

In this project, we adopted the following design standards:

- Data Pre-processing: We followed the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology for data pre-processing, which includes data cleaning, data transformation, and data reduction.
- Machine Learning: We used the scikit-learn library, which follows the API design standards for machine learning algorithms.

5.2 Coding Standards:

We followed the PEP 8 (Python Enhancement Proposal 8) coding standards for Python, which includes:

1. Naming Conventions: We used descriptive and concise names for variables, functions, and modules.
2. Code Organization: We organized the code into logical modules and functions, with each function performing a single task.
3. Indentation: We used four spaces for indentation to mark the beginning and end of control structures.
4. Function Length: We kept functions short and concise, with a maximum length of 20-30 lines.
5. Imports: Import one module per line, also it should be in alphabetical orders.

By following these guidelines we ensured that our code is readable, maintainable, and consistent with the Python community's best practices.

5.3 Testing Standards:

We followed the IEEE 829 (IEEE Standard for Software and System Test Documentation) standard for testing and verification, which includes:

- Test Planning: We developed a test plan that outlined the testing approach, test cases, and expected results.
- Test Case Development: We developed test cases that covered all the functional and non-functional requirements of the project.
- Test Data Management: We managed test data to ensure that it was relevant, accurate, and complete.
- Test Environment Setup: We set up a test environment that was identical to the production environment.
- Test Execution: We executed the tests and reported the results.
- Test Evaluation: We evaluated the test results and identified defects and issues.

By following the IEEE 829 standard, we were able to ensure that our testing and verification process was thorough, systematic, and effective in identifying defects and issues. This helped us to deliver a high-quality project that met the specified requirements and expectations.

Chapter 6

Conclusion & Future Scope

6.1 Conclusion:

In conclusion, our project aimed to develop a predictive model that can accurately forecast flight fares based on various factors such as airline, class, stops, departure time, arrival time, source city, destination city, and days left for departure. We employed a range of machine learning algorithms and techniques, including data pre-processing, feature engineering, model selection, and hyper-parameter tuning.

Our results showed that the Extra Trees Regressor model performed the best among all algorithms, with an adjusted R square of 78%,. We visualized the actual and predicted prices using line and scatter plots, which showed a close relationship between the two.

We followed standard design, coding, and testing standards to ensure the quality and reliability of our project. We used UML diagrams to model the system architecture and design, followed PEP 8 coding standards for Python, and adopted IEEE 829 testing standards for testing and verification.

6.2 Future Scope:

Our project has several potential avenues for future research and development:

1. Incorporating Additional Features: We can explore incorporating additional features such as weather data, holidays, and special events to improve the accuracy of our model.
2. Using Advanced Machine Learning Techniques: We can experiment with advanced machine learning techniques such as deep learning, transfer learning, and ensemble methods to improve the performance of our model.
3. Real-time Data Integration: We can integrate our model with real-time data sources such as APIs and web scraping to provide more accurate and up-to-date predictions.
4. Developing a Web Application: We can develop a web application that allows users to input their travel preferences and receive personalized flight fare predictions.
5. Expanding to Other Modes of Transportation: We can expand our model to predict prices for other modes of transportation such as buses, trains, and hotels.
6. Improving Model Interpretability: We can work on improving the interpretability of our model by using techniques such as feature importance and partial dependence plots.
7. Handling Imbalanced Data: We can explore techniques to handle imbalanced data, such as oversampling the minority class or using class weights.

By pursuing these avenues, we can further improve the accuracy and reliability of our model, and provide more value to users.

References

[1] Arjun, K.P., Rawat, T., Singh, R., Sreenarayanan, N.M. (2022). Flight Fare Prediction Using Machine Learning. In: Mehra, R., Meesad, P., Peddoju, S.K., Rai, D.S. (eds) Computational Intelligence and Smart Communication. ICCISC 2022. Communications in Computer and Information Science, vol 1672. Springer, Cham.

[2] K. Tziridis T. Kalampokas G.Papakostas and K. Diamantaras "Airfare price prediction using machine learning techniques" in European Signal Processing Conference (EUSIPCO), DOI: 10.23919/EUSIPCO .2017.8081365L. Li Y. Chen and Z. Li" Yawning detection for monitoring driver fatigue based on two cameras" Proc. 12th Int. IEEE Conf. Intel. Transp. Syst. pp. 1-6 Oct. 2009.

[3] Bhosale N, Gole P, Handore H, Lakde P, Arsalwad G. Flight Fare Prediction System Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology (IJRASET). 2022;10(1):931-8.

[4] Gupta S, Gupta N. Flight Fare Prediction Using Machine Learning.

[5] Subramanian RR, Murali MS, Deepak B, Deepak P, Reddy HN, Sudharsan RR. Airline fare prediction using machine learning algorithms. In2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT) 2022 Jan 20 (pp. 877-884). IEEE.

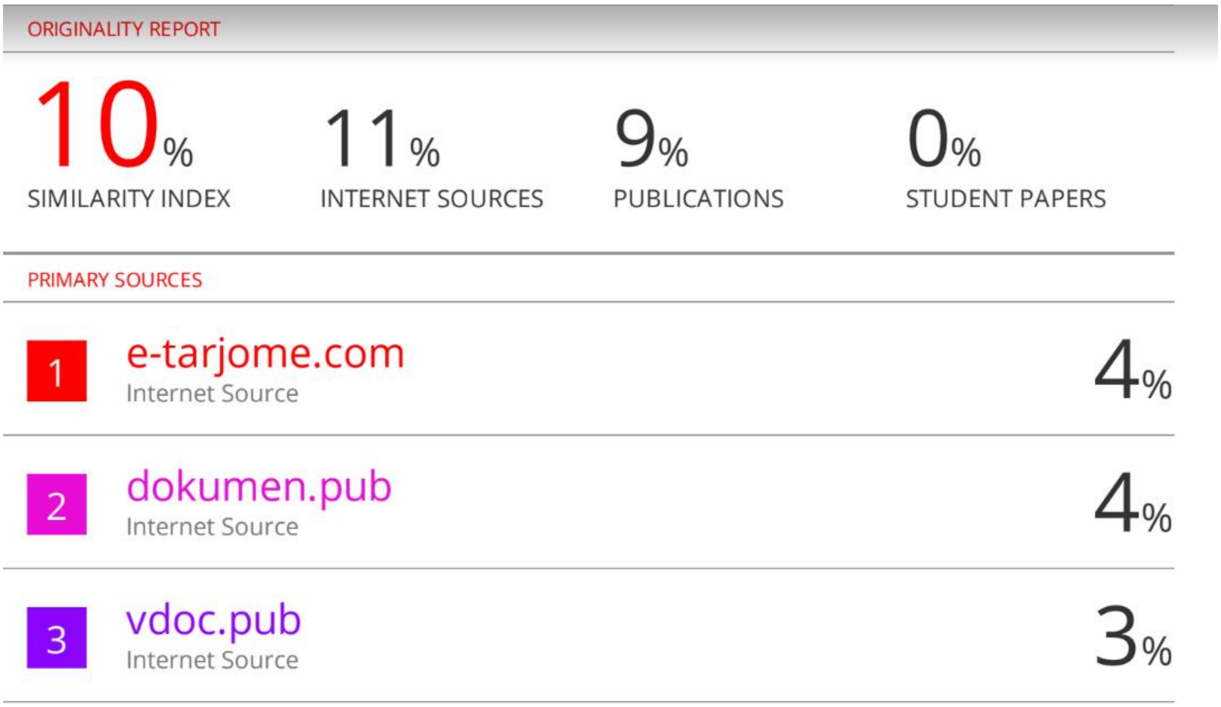
[6] Wang T, Pouyanfar S, Tian H, Tao Y, Alonso M, Luis S, Chen SC. A framework for airfare price prediction: a machine learning approach. In2019 IEEE 20th international conference on information reuse and integration for data science (IRI) 2019 Jul 30 (pp. 200-207). IEEE.

[7] Tuli M, Singh L, Tripathi S, Malik N. Prediction of Flight Fares Using Machine Learning. In2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence) 2023 Jan 19 (pp. 13-18). IEEE.

[8] Jayatkar K, Jagtap D, Dengale P, Satam A, Nivangune M. A Flight Fare Prediction Using Machine Learning.

[9] https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh

TURNITIN PLAGIARISM REPORT



ORIGINALITY REPORT

10% SIMILARITY INDEX

11% INTERNET SOURCES

9% PUBLICATIONS

0% STUDENT PAPERS

PRIMARY SOURCES

1	e-tarjome.com Internet Source	4%
2	dokumen.pub Internet Source	4%
3	vdoc.pub Internet Source	3%