Title:

A Comprehensive Study on Flight Price Prediction Using Multiple Machine Learning Algorithms.

Authors : *Atharva Wadhe, Archit Gajbhiye, Gautam Gogulwar, Ishika Ingole, Yash Ghuge*

Abstract

This paper explores the problem of flight fare prediction using multiple machine learning models to develop an accurate fare estimation system. The study presents a detailed analysis of a dataset comprising flight-related features such as duration, stops, and airline type. Six machine learning models—Linear Regression, Random Forest Regression, XGBoost Regression, K-Nearest Neighbors (KNN), and Light Gradient Boosting Machine — were evaluated. Performance metrics including R^2 score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were used to compare the models' effectiveness. The results demonstrate that XGBoost Regression outperforms other models with a significant reduction in RMSE, indicating its robustness in solving the flight fare prediction problem. This research highlights the importance of using advanced algorithms for accurate fare forecasting, providing valuable insights for the travel industry and laying the groundwork for future exploration in price prediction and demand forecasting.

Introduction

1.1 Background

In the last few decades, remarkable development in the field of machine learning (ML) has been a major factor in providing solutions to difficult problems faced in many sectors, including that of travel and especially aviation. In most cases with the introduction of online booking systems and dynamic pricing, predicting flight fares has – as it should – become a key issue in assisting consumers and even more service providers. Predicted fares, services optimally to the travel needs of people, help stave off losses and fully utilize the available resources for the airlines.

Nevertheless, while emphasis has been put on the term machine learning in the context of pricing strategies, the actual implementation of modern algorithms such as XGBoost, LightGBM, and other ensemble architectures to the task of forecasting ticket prices is still lacking. The main obstruction encountered when predicting fares is the incoming prices which tend to change under various circumstances such as demand, seasonality, and time to the flight. This research proposes a solution to this problem by exploring and comparing the effectiveness of modern machine learning techniques in predicting flight fares. Hence, with the utilization of these

elevated models, this study aims to address the shortcomings of current techniques and improve the prediction rate of fares in the domain of travel and air transport services.

1.2 Research Problem

With regard to ticket pricing predictions, the airline industry presents the greatest problem because of the multiple variables that oscillate the fare such as time of buying the ticket, season, prevailing demand by the airline among many other factors. This situation makes the prediction of fares difficult and manageable with a set of models that are able to render reliable predictions on the return fares.

Although a number of artificial intelligence models have been introduced many years back to solve the price prediction problems in various sectors, nowadays most of the academic researches and all of the industries do usages of first generation models or do not compare such advanced models at all. Such approaches are often unable to represent the interactions among various factors adequately, resulting in low-accuracy and low-generalization predictions. Because of this, firms and their consumers are not able to make the most optimal decisions concerning their pricing behavior.

The aforementioned issues will be dealt with by this study attesting and comparing the efficiency of superior ML algorithms such as XGBoost, LightGBM, Random Forest and so on to enhance the flight fare prediction accuracy.

1.3 Objectives

This research focuses on the following objectives:

1. To evaluate the effectiveness of multiple machine learning algorithms, including Linear Regression, Random Forest, XGBoost, KNN, LGBM, and SVM, for predicting flight fares.
2. To compare the performance of advanced models like XGBoost and Random Forest against simpler models such as Linear Regression and KNN in terms of accuracy and error metrics on a flight fare dataset.
3. To analyze the impact of key features such as departure time, travel duration, and airline choices in predicting flight prices.
4. To determine the best-performing model for flight fare prediction based on key metrics like R^2 score, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
5. To provide insights into the significance of each model for optimizing pricing strategies in the airline industry.

2. Related Work

Previous studies have explored various approaches for flight price prediction, each utilizing different machine learning techniques:

This research extends upon these prior works by employing a more diverse set of algorithms, including **ensemble methods** such as **Random Forest, Gradient Boosting, XGBoost, and CatBoost**. The goal is to improve not only predictive accuracy but also computational efficiency. The inclusion of **ensemble techniques** helps to address the **non-linear pricing trends** more effectively while maintaining manageable computational demands, thereby offering a more **balanced approach** to flight fare prediction.

This new research on airfare forecasting has attempted various machine learning models for higher accuracy and more solutions to the problem of forecasts of ticket price. As a matter of fact, one such report included eight state-of-the-art models that scored up to **88%** for an accuracy using a dataset of 1,814 flights. This variation between feature sets also is interesting, pointing to complex dependence relationships embedded in the airfare pricing. **Linear Regression** worked fairly well but was *outperformed* in price-volatile scenarios with non-linear dynamics. **Random Forest** performed reasonably well, but the influence of time of booking and seasonality made this kind of feature very difficult to handle, which is less probable for high accuracy in prediction. **Support Vector Machines** were strong, but it had major problems with fare change. Apart from that, **XGBoost** was proven to be a good solution but overfits strongly seasonal data frequently. High precision could be achieved on **ARTIFICIAL NEURAL NETWORKS** but barely performed due to sparse datasets with the presence of an *outlier fare*. Recurrent Neural Networks-based **LSTM** does excellently for bringing out strong time dependencies. However, it comes along with high computational overhead. The Hybrids and the combination of different algorithms also improved accuracy but it required heavy computations. Still, after all that updating in the machine learning field, fare prediction is improved but some difficulties came - the primary issues of model interpretability, feature complexity, and volatility of pricing data inherently.

---

3. Methodology

3.1 Dataset Description

The dataset includes 10,683 observation units and 11 predictor variables, including: – *Airline; Date_of_Journey; Source; Destination; Route; Departure_Time; Arrival_Time; Duration; Total_Stops; Additional_Info and Price.* Some relevant data considerations and actions are discussed below:

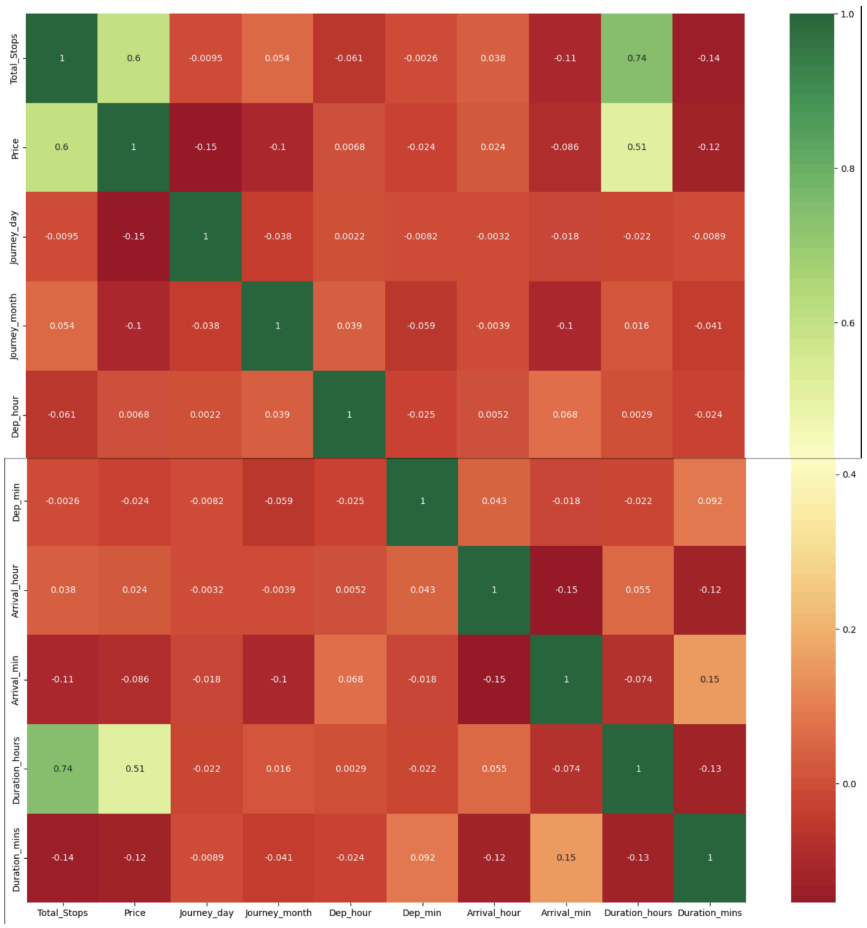

Fig 2 - Heatmap (Correlation of features)

Data Preprocessing is the preliminary phase of analysis where the dataset is cleaned and transformed into an acceptable format in the following ways:

- **Handling of Missing Values**: Rows with missing values were removed to maintain the quality of the dataset.
- **Normalization**: Features were transformed in order to reduce scales of differences between features which in turn leads to better performance of the model.
- **Feature Engineering**: New variables were created by parsing the Date_of_Journey for day and month.
  Departure_Time and Arrival_Time were made into hours and minutes for better clarification.
  The Duration feature was made into hours and minutes for better clarity.
- **Encoding Categorical Variables**: One-Hot Encoding was carried out on Airline, Source, and Destination to change these nominal variables into a form amenable to the application of the machine learning algorithms.

- **Dropping Redundant Feature**s: The Additional_Info property was eliminated from the data set because it contained more than 80 percent of “no info” records.

Furthermore, the Route and Total_Stops features were considered to have some interdependence and were later eliminated in order to streamline the analysis. This step of Preprocessing is important as it guarantees the neatness of the dataset and its suitability for training a model, thus ensuring its precision and reliability in making any prediction.

3.2 Machine Learning Models

This study implements several machine learning models to predict flight prices, including **Linear Regression, Random Forest Regression, XGBoost Regression, K-Nearest Neighbors (KNN), LightGBM (LGBM),** and **Support Vector Machine (SVM)**. Each of these methods offers unique advantages, which are briefly described below:

1. **Linear Regression**:

   A fundamental algorithm in regression tasks, Linear Regression establishes a linear relationship between the input features and the target variable. The model predicts the target by finding the best-fit line that minimizes the sum of the squared differences between observed and predicted values. Despite its simplicity, it serves as a strong baseline model.

2. **Random Forest Regression**:

   This ensemble method constructs multiple decision trees during training and outputs the mean prediction of individual trees. By averaging the results, Random Forest reduces the risk of overfitting, enhances robustness, and improves prediction accuracy. It can handle both categorical and numerical features, making it versatile for various datasets.

3. **XGBoost Regression**:

   An implementation of gradient boosted decision trees, XGBoost is designed for speed and performance. It optimizes the learning process through regularization techniques, which help prevent overfitting. XGBoost effectively handles missing values and captures complex patterns in the data, contributing to its status as one of the most powerful algorithms for structured data.

4. **K-Nearest Neighbors (KNN)**:

   KNN is a non-parametric, instance-based learning algorithm that makes predictions based on the k closest training examples in the feature space. The distance metric (usually Euclidean) measures the similarity between data points. KNN is simple and effective, particularly in datasets where similar observations are expected to yield similar outcomes.

5. **LightGBM (LGBM)**:

   A gradient boosting framework that uses tree-based learning algorithms, LightGBM is designed for speed and efficiency, especially with large datasets. It employs a histogram-based learning technique, which reduces memory usage and speeds up the training process. LGBM also provides better accuracy and faster training times compared to other boosting algorithms.

These models were selected to provide a comprehensive analysis of flight price prediction, utilizing both traditional and advanced machine learning techniques to determine the most effective approach for this specific task.

3.3 Model Evaluation

To confirm the universal applicability of the results across various parts of the dataset, model efficacy assessment was conducted using k-fold cross-validation. In particular, a well-known 5-fold cross-validation technique was used, whereby the dataset was split into 5 equal, stratified chunks. For every round, four folds were used for training the model, while the remaining one was utilized for evaluating the model and this process was repeated until all the folds had been utilized enhancing the evaluation of the model's performance.

The models were evaluated based on several performance metrics including:

- **R^2 Score**: This measure indicates what proportion of the target variable's variance may be accounted for by the independent variables, thus indicating the degree to which a model may be fit.
- **Mean Absolute Error (MAE)**: This measure provides the approximate prediction values in all cases, thus revealing the degree of accuracy for the model's estimate in absolute currency units.
- **Mean Squared Error (MSE)**: This measures the average squared errors which shows how far or close the predicted values are from the actual.

- **Root Mean Squared Error (RMSE)**: this again is the square root of the MSE, which measures the extent of deviation of predicted values from the actual value and is particularly helpful in giving a perspective on the extent of prediction distortion.

In the case of comparing models, the R^2 score was considered the best productivity measure of the flight fare price model as it was the best predictor of the price variation explained by the model. The evaluation served to provide the advantages and shortcomings of all approaches, resulting in XGBoost being chosen as the most competent algorithm for the task.

3.3 System Architecture

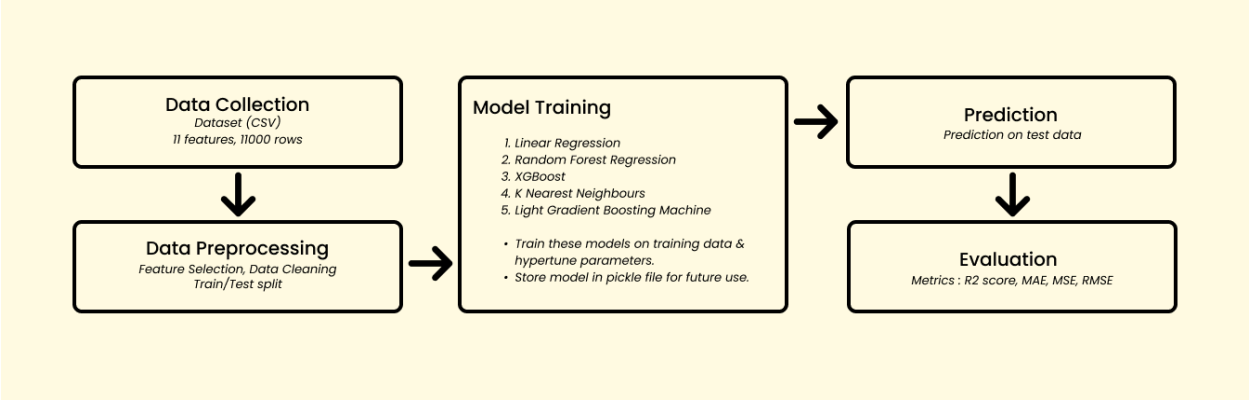

Fig 2 - System Architecture

---

4. Results and Discussion

4.1 Model Performance

The results of the analysis indicate that the **XGBoost algorithm** outperformed other models, achieving a **R^2 score of 0.846**, indicating that approximately **84.6%** of the variance in flight fare prices could be explained by the model. The performance metrics for the different models are summarized in Table 1.

Model	R^2 Score	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Linear Regression	0.620	1972.94	8202327.56	2863.97
Random Forest Regression	0.798	1260.43	3854031.99	1963.17
XGBoost Regression	0.846	1126.70	3321347.05	1822.46
K-Nearest Neighbors	0.793	1367.54	4462126.25	2112.37
Light Gradient Boosting Machine	0.826	1243.34	3755363.26	1937.88

Table 1 : Performance Metrics of all the models used in flight fare prediction

An overall comparative evaluation has been made regarding the models and it has been found that the **XGBoost model** is the optimal model for faring the flight expenses, achieving the best R^2 score, and the least MAE, MSE, and RMSE values. This shows that it is strong and effective in modeling the inherent complexities brought by the dataset, hence, it is the best option for this prediction task.

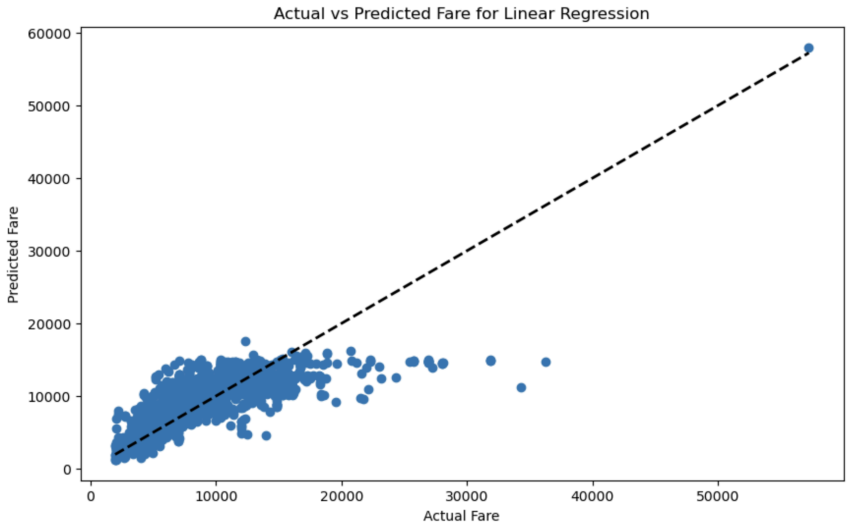

Fig. 3 - Actual vs Predicted Fare for Linear Regression

As seen in the **fig. 3**, it is clear that the model is proficient in the estimation of lower flight fares since the scatter points on the estimation versus actual graph are almost on the diagonal line which means that the cervical thrust is accurate. On the contrary, actual flight fares that are more than 20,000 seem to have been inaccurately estimated by the model as Bogolar's predictions are significantly less than the observed values. This shows that while the overriding model is effective in estimating the lower fares, there is need for improvement or alteration of the model in forecasting the higher fares.

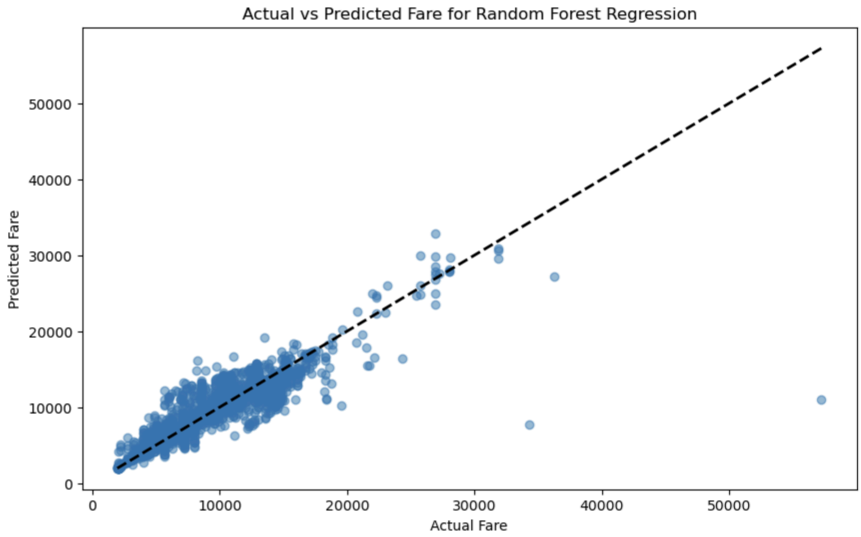

Fig. 4 - Actual vs Predicted Fare for Random Forest Regression

The Random Forest Regression model **(fig. 4)** is relatively accurate in terms of the prediction of flight fares as it could be observed that many points lie along the diagonal line, showing how closely the actual and predicted values agree with one another. For some reason, the model has a lot of trouble estimating higher fare predictions where the deviations from the ideal line become all too glaring. A few outliers that should not be there are predictors for improvement, but in general, the model is accurate in most of the situations.

The XGBoost fares plot **(fig. 5)** displays a reasonable fit for the lower and mid-range fares as the actual and XGBoost predicted fares fall within the same diagonal. But when the actual fare goes beyond a factor of 20,000, a few predictions go off the graph; some models undertake the lower end of the prediction for a few higher fare values. At the end of the day, the performance of XGBoost is quite satisfactory mostly for the fares below 30,000. But for higher fare predictions, XGBoost looks highly inaccurate for most of the participants and suggests room for improvement.

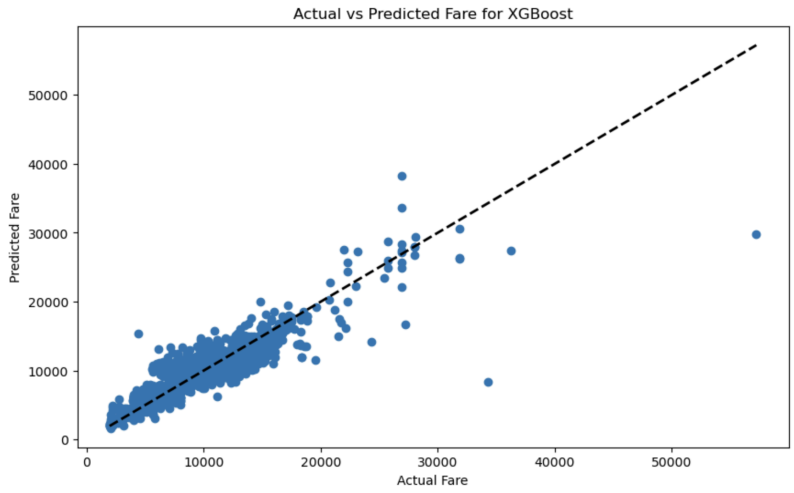

Fig. 5 - Actual vs Predicted Fare for XGBoost Regression

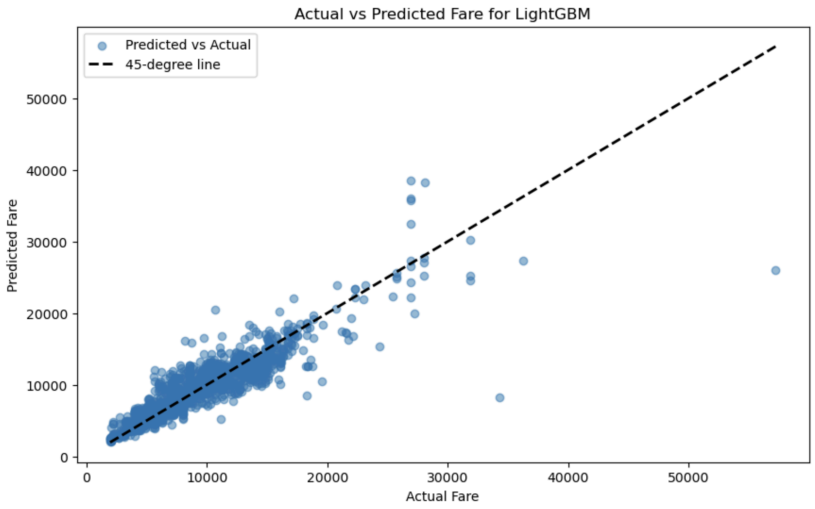

Fig. 6 - Actual vs Predicted Fare for LightGBM Regression

The LightGBM Regression model **(Fig. 6)** is evidently performing well as strength of fit indicated by the distribution of points around the 45-degree line most of the points are close to the diagonal line suggesting that predicted fares nearly agree with the actual fares. Nevertheless, there are a few clear outliers even at the higher fare predictions which implies that the model has slight variance when predicting these extreme cases. Generally, the use of this particular model corresponds to the real fares up to the reasonable outliners, which means that the model works properly in most situations.

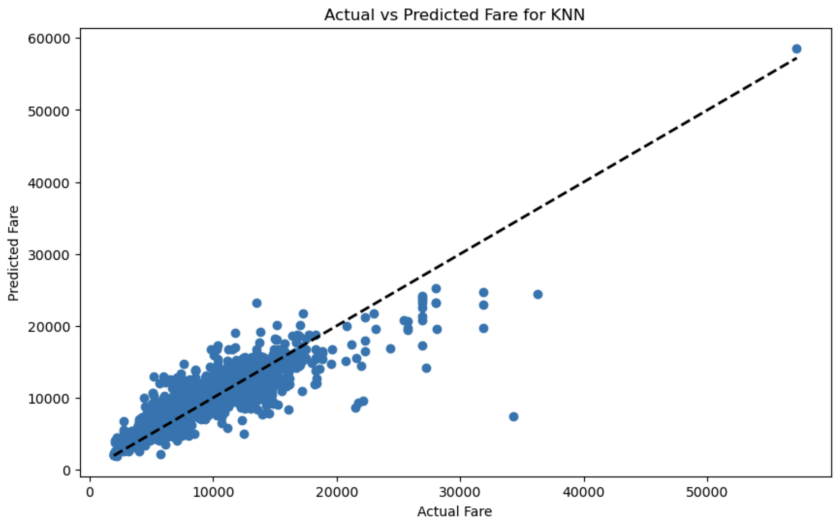

Fig. 7 - Actual vs Predicted Fare for KNN

According to the graph comparing actual fares and fares estimated using K-Nearest Neighbors modeling, **(Fig. 7)** the estimated fares were found to be comparable to the average fares in the lower fare ranges, […] as the dots appear on the lower diagonal of the line. As the actual fares went outside this range, the actual fare on prediction deviated from this average, with the trend being the prediction being to understate higher fare figures. KNN algorithm, K-Nearest Neighbors, is sufficient for making predictions on most low to mid-range fares; however, more work in extending the methods employed to higher fares is warranted as the extent for accuracy is noticeably hampered.

4.2 Discussion

The superior performance of the **XGBoost** model in predicting flight fares can be attributed to several key factors. XGBoost's ability to handle both continuous and categorical features effectively allows it to capture complex relationships within the data, which is crucial given the diversity of variables such as departure date, airline, flight duration, and price. Its implementation of gradient boosting enables it to iteratively improve its accuracy by minimizing errors at each step, leading to more precise predictions.

In contrast, models like **Linear Regression** and **K-Nearest Neighbors (KNN)** performed relatively poorly. This is likely because Linear Regression assumes linear relationships between features and the target variable, which may not capture the non-linear patterns in this dataset. KNN, on the other hand, might struggle with high-dimensional datasets and is sensitive to the choice of distance metric and neighborhood size, which can result in higher errors.

The **Random Forest** and **LightGBM** models performed reasonably well, with higher R^2 scores than Linear Regression and KNN, but still fell short of XGBoost. Both of these models excel at handling large datasets with numerous features and can effectively manage non-linear interactions between features. However, XGBoost's superior tuning capabilities and its more aggressive boosting strategy gave it an edge over these models.

Feature importance analysis further supports the robustness of XGBoost, highlighting that key features such as **departure time, total duration, and airline type** played a significant role in model prediction. This aligns with the initial objectives of this study, which aimed to evaluate the effectiveness of different machine learning algorithms in predicting flight fares. The results validate that **XGBoost** is the most suitable method for this task, providing more reliable and accurate predictions than the alternatives.

4.2 User Interface

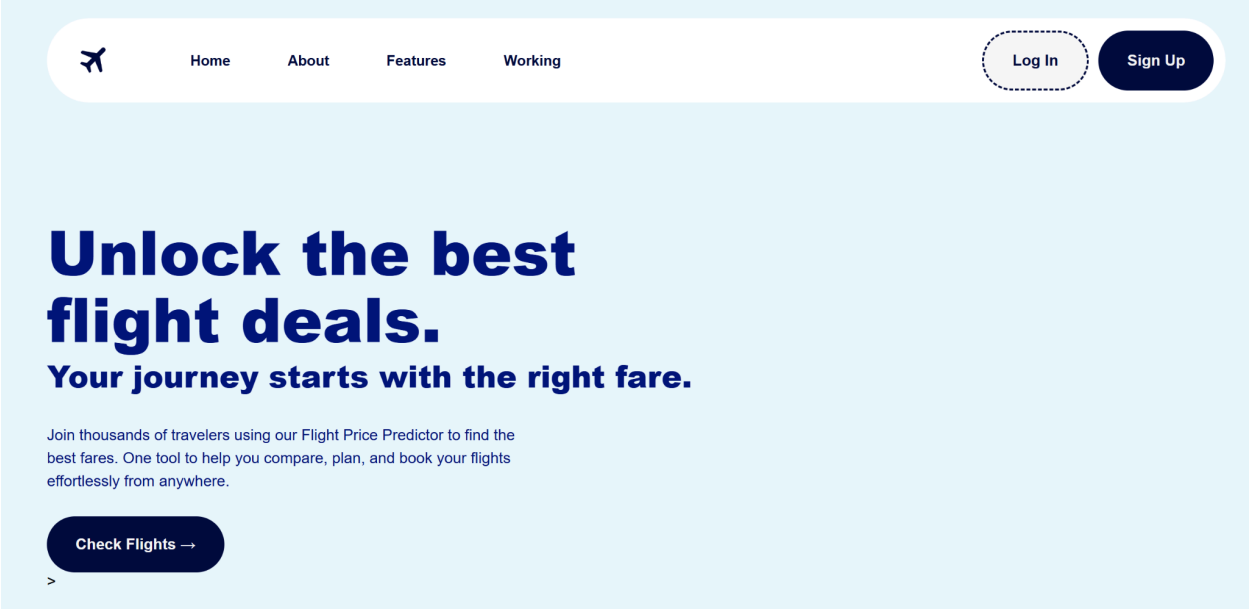

Fig 8 - Landing Page

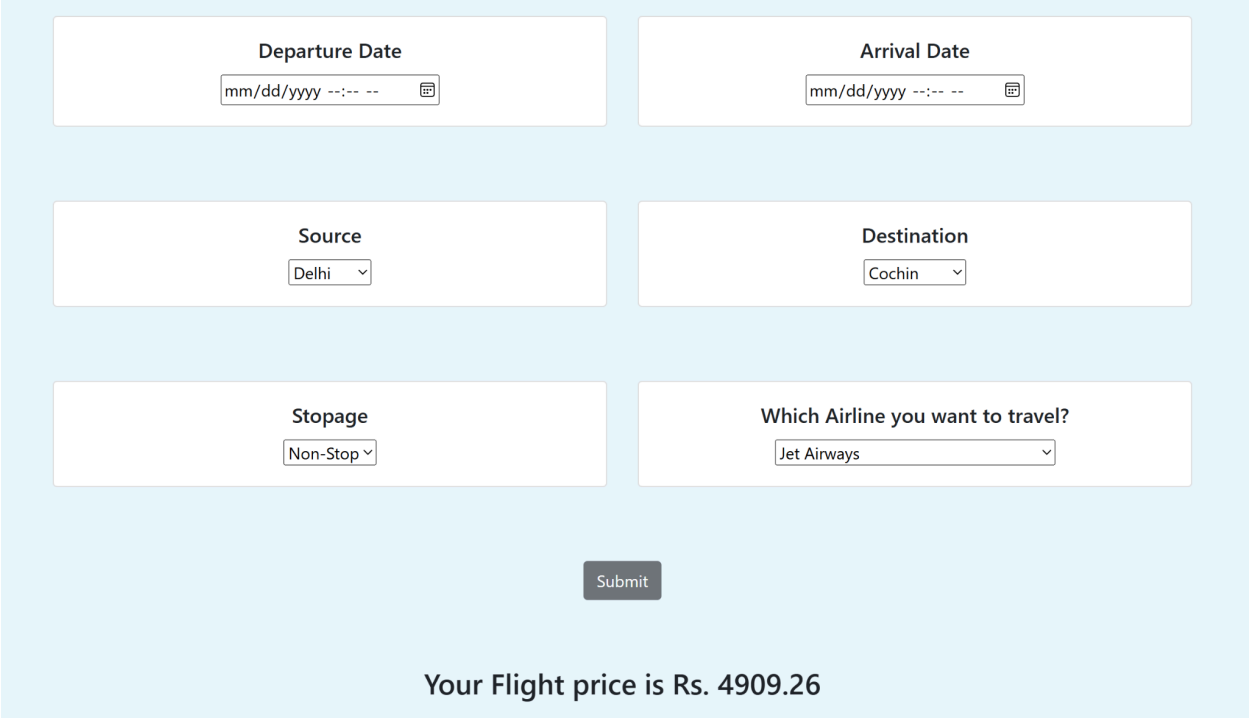

Fig 9 - Fare Prediction Page

---

5. Conclusion

This study shows that it is possible to build a model predicting flight fares with the use of XGBoost, which outperforms other methods commonly used, e.g., Linear Regression, Random Forest, or K-Nearest Neighbors. XGBoost was the most successful among all the models as it registered the highest R^2 and the lowest MAE, MSE, and RMSE values signifying its capability of understanding the different factors that determine flight fares.

The research asserts the significance of feature engineering techniques, for instance, extracting date, time and duration elements to enhance the performance of the created models. Yet, the fact that there were some limitations particularly the amount and diversity of the available data implies that future studies may delve into bigger datasets that may include actual flight information in real time or implement improved methods such as deep learning models. On top of that, more features such as weather information or fuel costs for instance can also enhance accuracy of the predictions.

Also, future work could extend to hyperparameter tuning using optimization approaches, such as Bayesian Optimization or Genetic Algorithms, which can lead to even more predictive accuracy improvement. Additionally, it would be desirable to extend the application of such predictive models to other fields, for instance, tourism, hotel pricing and transport.

---

6. References

[1] K. Tziridis, Th. Kalampokas, G. A. Papakostas, K. I. Diamantaras, "Airfare prices prediction using machine learning techniques," 2017 25th European Signal Processing Conference, IEEE, 2017.

[2] Rashid Nadeem, T. Sivakumar, "Flight Fare Forecasting: A Machine Learning Approach to Predict Ticket Prices," Proceedings of International Conference on Data Analytics and Insights, ICDAI 2023, Lecture Notes in Networks and Systems, vol. 727, Springer, 2023.

[3] Worku Abebe Degife, Bor-Shen Lin, "Deep-Learning-Powered GRU Model for Flight Ticket Fare Forecasting," Applied Sciences, vol. 13, no. 10, article 6032, 2023.

[4] Song S, Yang Y, Zhang Y (2019) Airfare prediction based on machine learning and expert feature engineering. In: 2019 IEEE international conference on service operations and logistics, and informatics (SOLI), pp 59–64.

[5] Gupta V, Singh R, Jain A (2018) Machine learning approach for predicting airfare prices. In: 2018 2nd International conference on computational systems and information technology for sustainable solution (CSITSS), pp 1–6.

[6] Chen J, Zhang Y (2018) A comparative study of machine learning algorithms for airfare prediction. Expert Syst Appl 101:335–345.

[7] Li Y, Zheng J (2020) Airfare prediction based on machine learning: a case study of Shanghai-Hong Kong route. J Adv Transp 2020:1–10.

[8] Reddy M, Kumar B (2020) Predicting airfare prices using machine learning algorithms. In: 2020 International conference on electrical, electronics, communication, computer, and optimization techniques (ICEECCOT), pp 626–631.

[9] Baral S, Khanal S, Adhikari S (2019) A hybrid machine learning model for predicting airfare prices. In: 2019 10th international conference on computing, communication and networking technologies (ICCCNT), pp 1–6.

[10] Alomari M, Alzubaidi M (2020) Airline ticket price prediction using machine learning. J Appl Eng Sci 18(1):47–53.

[11] Patil V, Pradhan P (2018) Airfare prediction system using machine learning algorithm. Int J Adv Res Comput Sci 9(2):79–84.

[12] W. Groves and M. Gini, "A regression model for predicting optimal purchase timing for airline tickets", Technical Report 11-025, 2011.

[13] G. A. Papakostas, K. I. Diamantaras and T. Papadimitriou, "Parallel pattern classification utilizing GPU-Based kernelized slackmin algorithm", Journal of Parallel and Distributed Computing, vol. 99, pp. 90-99, 2017.

[14] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola and V. Vapnik, "Support vector regression machines", Advances in neural information processing systems, vol. 9, pp. 155-161, 1997.

[15] S. B. Kotsiantis, "Decision trees: a recent overview", Artificial Intelligence Review, vol. 39, no. 4, pp. 261-283, 2013.